

赋能大数据教育

全国高校大数据教育教学经验谈

姚乐 朱启明 主编

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

未经许可，不得以任何方式复制或抄袭本书部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

赋能大数据教育：全国高校大数据教育教学经验谈/姚乐，朱启明主编. —北京：电子工业出版社，2018.3

ISBN 978-7-121-33766-6

I. ①赋… II. ①姚… ②朱… III. ①高等教育—教学经验—中国 IV. ①G649.2

中国版本图书馆 CIP 数据核字（2018）第 037943 号

策划编辑：缪晓红

责任编辑：董亚峰 特约编辑：刘广钦

印 刷：

装 订：

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本：720×1 000 1/16 印张：22 字数：350 千字

版 次：2018 年 3 月第 1 版

印 次：2018 年 3 月第 1 次印刷

定 价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254760。

编 委 会

主任委员

陈国良 中国科学院院士 全国高校大数据教育联盟理事长

副主任委员

卜佳俊 浙江大学计算机科学与技术学院常务副院长

孔繁之 济宁医学院医学信息工程学院院长

王兴伟 东北大学软件学院院长

王万良 浙江工业大学计算机科学与技术学院院长

王国胤 重庆邮电大学研究生院院长

孙名松 上海科技大学图书信息中心总工程师

陈立潮 太原科技大学计算机与科学学院院长

何宗耀 河南城建学院计算机与数据学院院长

李春生 东北石油大学计算机与信息技术学院院长

李劲华 青岛大学数据科学与软件工程学院教授、博士、副院长

李 鹏 哈尔滨理工大学软件学院副院长

李 涛 南京邮电大学计算机学院院长

汪 卫 复旦大学计算机科学技术学院副院长

张华平 北京理工大学计算机学院副教授

郝志峰 佛山科学技术学院校长

赵冬梅 河北师范大学信息技术学院院长

舒红平 成都信息工程大学软件工程学院院长

委员

- 马宏宾 北京理工大学自动化学院教授、博士生导师
- 方志军 上海工程技术大学电子电气学院院长、教授
- 王 鹏 西南民族大学计算机学院教授
- 王 伟 同济大学计算机科学与技术系副教授
- 王 鑫 中国传媒大学信息工程学院副教授
- 甘健侯 云南师范大学民族教育信息化教育部重点实验室教授、常务副主任
- 安小米 中国人民大学信息资源管理学院教授
- 刘 刚 北京邮电大学世纪学院通信与信息工程系副主任、副教授
- 任 鸣 浙江旅游职业学院教授、高级工程师
- 许 伟 中国人民大学信息学院副教授
- 陈 禹 中国人民大学信息学院教授、博士生导师
- 陈学斌 华北理工大学数据科学实验中心主任、教授
- 李 辉 中国农业大学信息与电气工程学院博士、副教授
- 张德富 厦门大学信息科学与技术学院计算机系副主任
- 张 晖 西南科技大学教授
- 张祖平 中南大学信息院教授
- 金义富 岭南师范学院网络与信息技术中心主任、教授
- 林子雨 厦门大学信息科学与技术学院助理教授
- 周丰丰 吉林大学计算机科学与技术学院教授、博士生导师
- 洪文兴 厦门大学自动化系副教授
- 胡福文 北方工业大学机械与材料工程学院副教授
- 胡学钢 合肥工业大学计算机与信息学院教授
- 高峻峻 上海大学悉尼工商学院教授

- 郭景峰 燕山大学信息科学与工程学院计算机系教授、博士生导师
- 倪建成 曲阜师范大学软件学院教授
- 夏 天 中国人民大学信息资源管理学院副教授
- 曹淑艳 对外经济贸易大学信息学院教授、信息化处副处长
- 黄润才 上海工程技术大学电子电气工程学院副教授
- 萧 冰 上海交通大学媒体与设计学院副教授
- 朝乐门 中国人民大学信息资源管理学院副教授
- 董小英 北京大学光华管理学院副教授、博导
- 董付国 山东工商学院计算机科学与技术学院副教授
- 傅湘玲 北京邮电大学软件学院副教授
- 谭维智 曲阜师范大学教授
- 秦松疆 全国高校大数据教育联盟副理事长 章鱼大数据
CEO
- 姚 乐 全国高校大数据教育联盟秘书长 CIO 时代学院
院长
- 朱启明 全国高校大数据教育联盟副秘书长 CIO 时代 APP
总编
- 张士运 全国高校大数据教育联盟副秘书长 章鱼大数据
COO
- 刘娜艺 全国高校大数据教育联盟实验室主任
- 王甲佳 CIO 时代学院同学会秘书长
- 鲁四海 北大CIIM—泸州电子商务大数据开放实验室副主任
- 侯丽敏 CIO 时代 APP 记者编辑 CIO 时代品牌研究院研究员
- 孔 文 CIO 时代 APP 记者编辑 CIO 时代品牌研究院商务
总监

序 言

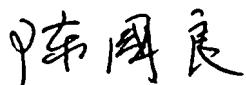
教育关乎国计民生，教育问题异常复杂。大数据在重塑教育教学方面具有无限的潜能。那么，大数据怎样才能有效驱动教育变革？大数据教学有哪些特点？大数据进入教育领域将面临哪些困难与挑战？近期，由 CIO 时代 APP、全国高校大数据教育联盟主编的《赋能大数据教育：全国高校大数据教育教学经验谈》一书，对这些问题进行了很好地诠释。

《赋能大数据教育：全国高校大数据教育教学经验谈》这本书旨在为教育界、产业界及社会各界人士打开一扇了解“大数据教育”的窗户，本书结合全国 42 所高校 51 位院长、副院长、教授、副教授有关教育大数据的学术讲座进行内容编辑，共分 6 篇，分别为教育篇、人才篇、产业篇、技术篇、应用篇、未来篇。虽然篇幅不长，但本书编者通过艰苦的研究，精心编辑了最新的、有重要价值的大数据教育教学领域的学术研究成果、教育行业的重要数据分析等，深入浅出地介绍了什么是大数据教育、大数据教育的价值、大数据教学方法、大数据挖掘、大数据应用、大数据可视化、大数据安全、大数据前瞻等，从不同的角度为读者展示了一幅浩瀚的高校教育大数据蓝图。面对汹涌澎湃的大数据，无论对于高校大数据教育工作者还是普通公众，这本书无疑都具有重要的参考价值。

当前，中国的高等院校培养了大量的计算机、软件、统计分析、数学应用、社会管理等专业人才，为中国的大数据战略进行了很好的人才储备，但从发展来看，这些储备的人才还远远满足不了社会需要。随着全国高校陆续开办“数据科学与大数据技术”“大数据技术与应用”等大数据专业，未来，将会有更多的、更专业的大数据人才从高校走出来，源源不断地进入市场为社会服务。我们应在大数据时代抓住历史机遇，与时俱进，为实现中国经济结构转型和中华民族伟大复兴贡献力量。

与其他行业相比，教育界对大数据的广泛接纳还是近期的事。可喜的是，我们看到大数据正在走进教育的领地、走进学校的大门、走进教师和学

生的生活。可以预期的是，一个属于教育的大数据时代即将到来，它不仅影响学校内部治理的改革，而且会驱动整个教育领域的变革。但从整体来看，目前大数据在教育决策、教学过程中的运用还处于摸索和起步阶段，大数据人才培养体系仍需要不断完善。这些遗留工作，仍需要产、学、研各界相互努力，协同发展，让我们携起手来为我国教育大数据事业奋斗。



中国科学院 院士

全国高校大数据教育联盟 理事长

2018.1

前言

科技强国，教育兴邦。赋能大数据教育，培养大数据人才，是实现强国兴邦的重要途径之一。2015 年我国正式启动国家大数据战略，涵盖教育、医疗等多个领域。其中，大数据教育处在优先发展位置，这是国家教育事业全方位变革与创新发展的必然要求。然而，当前我国教育大数据研究与实践工作整体还处于起步探索阶段，国外也没有成熟的经验和模式可以借鉴。

基于此，2017 年由全国高校大数据教育联盟推荐，邀请来自全国 42 所高校 51 位计算机学院、软件学院、统计学院、数学学院、大数据学院的院长、副院长、教授、副教授，在 CIO 时代 APP 微课堂平台上，担纲主讲了大数据教育教学及新一代 IT 应用主题内容。北京大学、复旦大学、浙江大学、中国人民大学、中国农业大学、北京理工大学、同济大学、厦门大学、东北大学、对外经贸大学、合肥工业大学、浙江工业大学、重庆邮电大学、太原科技大学等一批高校领导、专业教师，结合大数据教学工作内容进行分享，旨在为高校大数据教育事业献计献策，提供解决方案。

他们是时代弄潮儿，是全国高校大数据教育工作者的缩影，是第一批“吃螃蟹”的人，为高校大数据教育事业发展积累了丰富的理论和实践经验；他们是宣传队，向全国高校和社会传播大数据教育教学领域的先进思想、先进方法和先进技术；他们是“播种机”，向全国高校和社会撒下大数据教育的种子，普及大数据知识、传播大数据文化、汇聚大数据力量。

为了破解大数据教育教学工作难题，他们前期投入了大量时间和精力准备讲座内容。有的老师提前 3 个月收集资料，反复修改讲座内容；有的老师在出差途中的高铁上、飞机上构思讲座选题……总之，每一场微讲座，每一次主题分享，有观点，有亮点，有干货。50 多场 CIO 时代 APP 微讲座，基本勾勒出高校大数据教育教学的知识体系，为我国高校大数据教育事业的发展提供了理论基础和决策依据，填补了教育大数据市场的一项空白。

为了记录、传播、收藏这些宝贵的知识点，应广大读者和网友要求，CIO

时代 APP 编辑经过前期精心策划、重新梳理讲座内容并集结成书——命名《赋能大数据教育：全国高校大数据教育教学经验谈》，此书为大数据教育主管部门领导、高校师生、企业大数据应用开发人员、大数据中介培训服务机构及大数据行业从业者提供决策参考。

《赋能大数据教育：全国高校大数据教育教学经验谈》是集体智慧的结晶，书中公开出现的大数据教育理论、观点、案例在业界尚属首次。全书内容共分 6 篇。教育篇系统介绍了当前本科院校数据科学与大数据技术专业申报过程、存在的问题及基本解决思路、研究并分析了高校大数据实验室建设及大数据课程体系设计与规划路径；人才篇重点分析了“互联网+”时代，高校大数据人才培养的思路、方法，以及需要注意的问题；产业篇强调了产教融合，学以致用，突出了大数据与企业管理、企业转型的商业辩证关系；技术篇阐述高校大数据教育围绕技术开展实践教学，以数据挖掘、数据可视化、人工智能、3D 打印等为代表的新一代 IT 为基础，面向行业如何实现商业变革；应用篇围绕行业，重点分析了教育、工业、金融、健康、能源、媒体、旅游等行业大数据应用成效；未来篇介绍大数据即大智慧，前瞻大数据安全、技术、应用带来的战略思考、机会与挑战。本书内容朴实无华，既有对高校大数据教育专业的深刻认识、大数据课程体系总体规划、大数据实践教学的案例分析，也有对目前高校教育改革问题的探讨与思考。期望对教育大数据行业从业者有所启发，有所帮助。

同时，本书在编写过程中得到了全国高校大数据教育联盟成员单位领导、老师，以及广大同行专家、企业家的大力指导和帮助，特别是章鱼大数据的支持，在此一并表示感谢。在国内，教育大数据刚刚起步，正处在改革创新过程中。截稿前，全国已经有 35 所高校开办“数据科学与大数据技术”本科专业，2017 年又有 263 所高校正在申报大数据专业，但这个数字占全国高校比例仍然很小。因此，有关高校大数据教育的理论研究与实践体系仍不成熟，再加上 CIO 时代 APP 编辑能力和经验所限，不妥或疏漏之处在所难免，请各位同仁多提宝贵意见。

大数据专业是一门新学科，交叉性强，跨度大且复杂，完善一门学科建设需要很长时间。基于大数据的个性化教学、科学化评价、精细化管理、智能化决策、精准化科研等，将对促进教育公平、提高教育质量、培养创新人

前 言

才具有不可估量的作用。数据驱动教育创新、数据驱动教育变革已成为不可更改的趋势。大数据教育事业任重道远，前途光明，期望能有更多的同行加盟进来，共同为我国教育大数据事业发展贡献力量。

姚 乐

全国高校大数据教育联盟 秘书长

CIO 时代学院 院 长

2017 年 12 月

目 录

教育篇

从大数据到数据科学

“数据科学与大数据技术”专业课程体系与教学环节探讨	曹淑艳 (3)
为什么《数据科学》是现代人才的“必修课”	朝乐门 (9)
数据科学与大数据技术——新工科、新探索	方志军 (13)
当“数据科学”遇上“自由博雅”	王 伟 (19)
数据科学与大数据技术专业申报与建设	张祖平 (26)
关于高校大数据教学若干关键问题的探讨	林子雨 (32)
运用大数据技术深化教育信息资源应用	舒红平 (40)
大数据在高校智慧校园中的应用	孙名松 (46)
区块链+教育的需求分析与技术框架	金义富 (55)
本科大数据实验平台及资源建设等的思考与探索	李 辉 (59)
民族教育信息化建设探索与研究	甘健侯 (64)
大数据在教育中的应用及限度	谭维智 (71)
大数据背景下的“新工科”培养模式——以软件工程为例	倪建成 (77)

人才篇

大数据人才培养之道

“互联网+”时代创新人才培养模式的思考	郝志峰 (85)
大数据应用创新及人才培养探讨	何宗耀 (91)
大数据技术人才培养需要跨越的障碍	胡学钢 (97)
大数据研究要注意的两个问题	陈 禹 (101)
对大数据专业人才培养的几点思考	黄润才 (105)

产业篇

新时代下的互联网产业变革

“互联网+”背景下的企业转型与变革·····	董小英 (113)
大数据与企业管理决策·····	傅湘玲 (117)
大数据产业中的协同创新——技术、应用与新业态的区域实践 ·····	洪文兴 (124)

技术篇

互联网创新驱动，推动技术研发

大数据技术与产业中的几个关键问题商榷·····	陈立潮 (133)
Python 编程要点·····	董付国 (139)
数据可视化·····	刘 刚 (146)
大数据时代的数据挖掘·····	李 涛 (150)
大数据时代的人工智能·····	王国胤 (158)
3D 打印格局的大视野认知·····	胡福文 (162)
人工智能——奇点还是支点·····	马宏宾 (168)
多尺度量子谐振子优化算法·····	王 鹏 (185)
深度学习的最新进展·····	王万良 (191)
大数据时代的编程语言·····	夏 天 (196)
人工智能时代的“盲点”——信息无障碍·····	卜佳俊 (203)

应用篇

融合先行，释放价值

工业/物联网大数据·····	汪 卫 (213)
大数据语义分析与应用实践·····	张华平 (218)
油田大数据应用探索·····	李春生 (228)

广播电视个性化节目推荐系统·····	王 鑫 (233)
大数据与健康·····	孔繁之 (240)
大数据+旅游——由旅游供给侧需求导出·····	任 鸣 (246)
金融大数据分析与应用·····	许 伟 (250)
健康大数据的降维问题·····	周丰丰 (255)
基于认知心理学的大数据可视化设计研究·····	萧 冰 (260)
大数据与生物信息学的应用研究与实践·····	李劲华 (263)
大数据治理规则体系构建研究·····	安小米 (268)
社会网络中顶点相似性度量方法研究与应用·····	郭景峰 (284)
10 大怪象——国外供应链计划和决策类软件在中国的运用·····	高峻峻 (297)
大数据和人工智能在高校舆情处理中的应用·····	张 晖 (303)
物联网与农业大数据·····	陈学斌 (308)

未来篇

大数据带来的机遇与挑战

大数据时代：机遇、挑战与思考·····	王兴伟 (315)
互联网+大数据=大智慧·····	张德富 (320)
在自由与控制之间达至创新·····	李 鹏 (326)
大数据安全机遇与挑战·····	赵冬梅 (329)
后记·····	(334)



教育篇

从大数据到数据科学



“数据科学与大数据技术”专业课程体系 与教学环节探讨

对外经济贸易大学信息学院教授 曹淑艳

一、大数据专业的认识

（一）大数据专业设置的理由和基础探讨

第一，国家的重视。我国十分重视大数据产业的战略意义、大数据资源对社会发展的作用，具体表现为国务院于 2015 年出台了《促进大数据发展行动的纲要》，并将实施国家大数据战略、推进数据资源开发共享纳入“十三五”期间规划建设的重要目标。从目前来看，国内的大数据产业发展已基本具备一定规模，正有待于形成产业界的共识。从国外情况来看，尤其是美国市场，麦肯锡在报告中预测，2018 年美国大数据人才和高级数据分析专家缺口达到 19 万人，美国企业还需要 150 万能提出正确问题、运用大数据分析结果的相关管理人才。同时，85%的 500 强企业已经或正在筹划推出大数据项目，未来几年这些企业在大数据的投资将上涨 36%。同时，大数据产业的发展也需要大数据人才的支撑。

第二，大数据相关专业的发展情况。国内 2015 年申报新专业，2016 年 3 月教育部批准三所高校设立本科专业，2016 年又有 38 所高校申报在目录外本科专业，其中既有 985、211 院校，也有地方院校，自从 2016 年 3 月开始首批设立专业后，全国上下的高校均开始重视起来。可以看出，大数据相关专业发展情况近几年在国内的势头较为迅猛。

第三，高校陆续进行了开设大数据专业的基础工作。如高校大数据联盟开办的相关会议，高校也加强了师资团队的培训与建设，探讨实践教学的建设，组建了大数据专业建设小组，与大数据产业协会的联系日益紧密，参与

了大数据人才培养论坛，探讨复合型人才就业前景，等等。这些工作都是目前全国高校在开展的。

第四，学科逐渐交叉。目前很多学科在进行学科建设时也在向大数据方向发展，如数学、统计、信息管理、管理科学与工程等都在向大数据专业靠拢。因此，高校在进行申报时将大数据归到信息科学与技术学科下，其实它是一门交叉学科，应该是一个融合，用来解决问题时是明确的，要通过数据分析和挖掘来获得价值。从目前来看，与2000年初设立电子商务专业相似，如今电子商务已成为管理学门类独立的学科，但有时也无法进行准确的归类。

（二）大数据专业和相近专业的区分和联系

大数据专业和相近专业的区分和联系主要体现在计算机科学技术、统计学、信息管理、数学四个学科上。

与计算机专业的区别和联系。第一，一名经验丰富的软件工程师转型到数据科学领域是很方便的，因为大量的大数据工作都涉及软件工程方面的知识，不仅包括设计健全的系统，也包括简单的软件。拥有这种计算机专业的学术背景，可能会快速完成实验任务。因此，大数据科学应用专业的技术支撑主要来源于计算机科学；第二，大数据专业具有前所未有的复合型特征；第三，以大数据为核心的研究对象强调学生对专业领域数据的理解能力，体现了技术为数据服务的思想，是复合型人才。二者既有联系又有区别，联系是计算机技术支撑着大数据专业。

与统计学专业的区别。统计学建模使用的数据是结构化数据，大数据的来源是多元、异构的，因此，其中会有一些区别。大数据重视非结构化、半结构化数据的处理，如今大数据强调处理技术平台、获取存储、处理和展示各个环节与计算机深度融合，统计学的方法是大数据环节进行数据分析必不可少的，联系很紧密，区别也是很清晰的。

与信息管理专业的区别。主要体现在看待数据和信息的角度。信息管理主要强调在理解数据和业务流程的基础上，通过科学的分析和设计方法来实现管理信息系统，强调利用计算机技术借助改造升级原有的业务系统，如学校中的人事管理系统、学生管理系统等，而大数据相关理论和技术管理系统侧重对数据本身的洞察和理解，相对而言独立于原有的业务系统，更专注于

海量、复杂、多元数据的深度分析处理能力，更依赖于大数据处理平台和技术，更好地支撑了物联网、移动互联网平台的发展。

与数学的关系。数学知识是为数据科学的数据分析所有研究领域打下坚实基础的学科，因此是做基础支撑。大数据是在数学的基础上进行应用分析。

二、课程体系

新设专业探索前行。这是一个新的专业，有关人才培养方案与课程体系难以成熟，难以有完整的课程体系，需要不断去探索。如现阶段的信息管理专业是在 1978 年设立的，电子商务专业是 2000 年初确定的，都是需要一个长时间的探索过程的。我们可以围绕大数据产业的核心要素，根据各个学校的特色进行大数据课程核心体系设计，设计过程中可将统计学的一些定量模型和方法纳入考虑范围。

从大数据产业的核心要素进行分析。以往在申报专业之初是依据数据价值的提升路径来进行课程体系的初步设计的，例如从数据的获取、存储、清洗开始，形成结构化数据、利用统计学方法进行统计分析，其结果利用可视化技术进行展示，进而支持行业应用，这是原来的方法。孙鑫则是从四个面对大数据产业的核心要素进行划分（见图 1），是值得借鉴的。其中，核心要素分为四个层次，是根据数据价值的提升路径及 IT 领域的产品布局来设立的。四个层次分别为数据资源层、基础能力层、分析展示层及应用层。数据资源的功能主要是负责原始数据的供给与交换，是数据资产作为生产要素的直接表现，根据数据来源不同可以细分为数据资源提供者和数据交易平台两种角色。数据基础能力部分是负责与数据生产加工相关的基础设施与技术要素的供应，为数据加工和价值提升提供生产工具，包括数据存储、数据处理和数据库等多个角色。数据分析和展示部分负责数据隐含价值的挖掘、数据关联分析与可视化展现，这是智力要素在数据价值中的集中体现，包括传统意义上的人工智能、可视化和通用数据分析工具、面向非结构化数据提供的语音图像等媒体识别服务。而数据应用部分是根据数据分析和加工的结果面向电商、金融、交通等细分行业提供精准营销、信用评估、初期引导等企业或公众的服务。

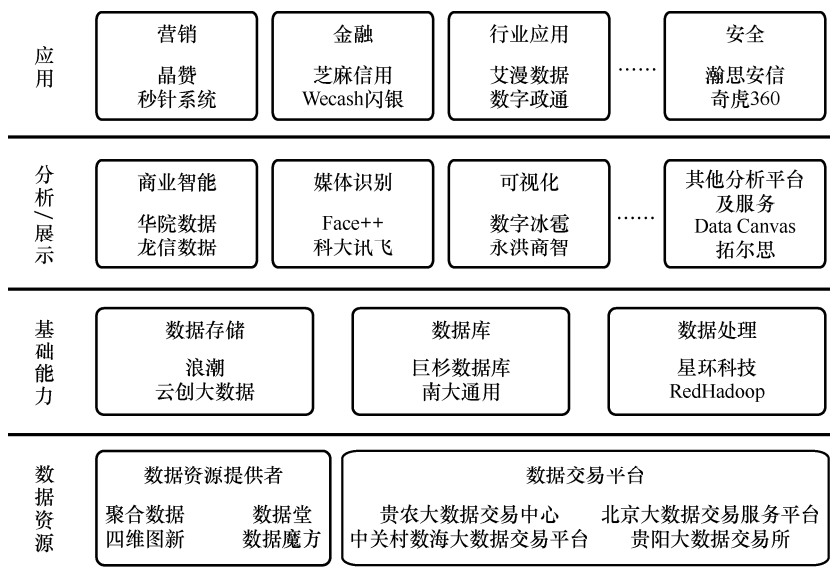


图1 大数据产业核心要素^①

围绕大数据产业核心要素的四个层次，可以看出在数据资源层和基础能力层主要还是与传统计算机科学与技术专业的核心课程设置相关联的。在数据资源提供和数据交易平台还涉及信息资源管理、外部技术、软件体系结构等。在数据基础能力层主要提供的是数据存储、数据处理和数据库，这些主要体现在计算机软件技术，如各种程序设计语言，体现在大数据上的是 R 语言、程序设计、数据结构等。在数据分析和展示时，体现在人工智能、统计学应用、数据挖掘、机器学习、可视化和通用数据分析、可视化和媒体识别服务等。主要的课程涉及分布式系统原理、人工智能基础、应用统计、多元统计分析、机器学习与数据挖掘、数据可视化、云计算及相关的大数据分析。在数据应用层建议结合高校的特色来进行设计，以对外经济贸易大学为例，将数据应用与专业特色、学校背景相结合，申报时分为以下三个方向：贸易金融大数据、网络营销大数据、电子商务大数据。如果想做贸易金融大数据，需在课程中包括国际贸易背景、财务会计概论、供应链管理、风险管理学；要想做网络营销大数据，要有计算广告学、搜索引擎优化、网络营销等课程

^① 孙鑫，中国信息通信研究院政策与经济研究所，我国大数据产业发展态势如何？

背景：要做电子商务大数据，可能围绕电子商务概论、电子商务应用基础、电子商务系统分析与设计、个性化推荐理论和实践来进行课程体系设计。从大数据产业的四个核心层次来看，进行课程体系梳理时可以看出，数据资源与数据基础能力依托于传统的计算机技术多一些，在数据分析和展示层是在统计学支撑下与 IT 相结合有所创新，在数据应用层取决于各个学校的办学特色，在应用层方面依靠高校的专业背景来进行相应的课程设计。

进行课程体系设计的目的是人才培养，学生未来的就业应该是沿着大数据分析师的职业方向走，接着可能会是大数据分析行业专家，这是业务层；技术层是向大数据架构师方向走；管理层是向大数据分析总监方向走，再向前走是首席数据官。推荐以下三个就业领域：贸易金融方向的大数据分析师可以在供应链融资公司、P2P 信贷征信平台、商业银行等找到自己的位置；网络营销方向的大数据分析师可以互联网广告、O2O 营销公司、大型网络媒体企业就职；电子商务方向的大数据分析师应沿循电子商务公司、第三方支付公司、电子商务流公司来做。因此，专业的课程体系设计应与学生的就业领域和职业生涯发展相匹配，于是可能会梳理出一些核心课程，如多元统计分析、统计分析软件与应用、统计学习理论与基础、人工智能、商务智能、数据科学导论、数据挖掘、数据可视化、模式识别、云计算学习中心、大数据分析等，都应作为课程体系中的核心课程。

三、实践性教学环节与主要实验

大数据这一学科实践性很强，学科交叉线也很强。这里将实践性教学环节按实验性质分为三个层面。

第一层是技术基础类的实验。包含数据库、程序设计、云计算平台的建设与开发等。

第二层是数据分析及展示类的实验。包括机器学习、大数据分析 with 处理、大数据可视化。

第三层是综合应用类实验。学生学到最后时，应有一门综合性的课程，像对外经济贸易大学这样的财经背景学校应该设有“金融大数据综合实验”“营销大数据综合实验”或“物流大数据综合实验”等课程，这种综合类应用

实验建议与相关的大数据公司建立产学合作关系，进行人才联合培养。另外，大数据联盟平台也提供了一些机会。

的确，大数据专业是很新的。没有十分完善的专业课程体系，只能在探索中前进。建设新专业任重而道远，最后的实践可能还会遇到问题，需要不断进行修正。但最后的目标是一致的，即如何将大数据这一新专业建立起来，为企业、市场提供人才。

作者简介

曹淑艳：对外经济贸易大学信息学院教授、信息化处副处长。从事计算机教学及信息系统开发研究工作三十余年，本科专业是计算机软件，获管理学硕士学位，金融学博士。参与了教育部高校文科计算机基础教学指导委员会《大学计算机教学基本要求》的编写工作，教育部大学计算机基础教学指导委员会文科分指导委员会副主任委员，全国高校计算机基础教育研究会文科专业委员会副主任委员，全国高校计算机基础教育研究会财经专业委员会课程建设委员会副主任委员。

为什么《数据科学》是现代人才的“必修课”

中国人民大学信息资源管理学院副教授 朝乐门

“数据科学”是大数据背后的科学。大数据热给我们带来的是各个学科领域所面对的“数据”变了，导致人们对数据的“认识”也发生了改变。当然，这还不是问题的关键，问题的关键在于，大数据这场“风暴”过后，它会留下《数据科学》。至于“数据科学”的规范定义，本文不做展开讨论。如果感兴趣，可以查阅作者撰写的两本书《数据科学》和《数据科学理论与实践》，书中给出了规范定义方法。

现在，很多人都在纠结于大数据的这个“大”字之上，都在试图诠释现代社会的“数据规模”有“多大”，其实这是一种“曲解”。所谓的“大”是相对概念。人类历史上，每过一段时间，人们都会觉得信息量“大”，“大”到“快要爆炸”。例如，一百年前的科学家也曾感到当时的信息已经“爆炸”，他们觉得学术论文一下子“多得不得了”，都看不过来，开始要求写“摘要”。从现代人的眼界看，一百年前的所谓“信息爆炸”不算什么。同样，如今所谓的“大数据”对于一百年后的人们来说可能也不算什么。

“大数据”的“奥妙”不在于其“大小”，而在于“我们所面对的数据已发生了改变”。近年来，随着“云物移大智”等新技术的普及，我们获得、存储和处理数据的能力提升了；结果是，我们所面对的“数据”变了；更重要的是，传统知识，如各领域中的传统理念、理论、方法、技术、工具等无法处理“这种变化了的新数据”；所以各学科需要重新认识“数据”，需要在认识论和方法论层次上重写自己所在学科领域的“知识”。

如果仔细观察，会发现一个很奇怪的现象——现在几乎所有的领域都在谈“大数据”，但是每个领域对“大数据”的理解都不同。每个领域都认为自己做的是“真正的大数据”，总是怀疑另一个领域所说的“大数据”并不是“真正的大数据”。大家不要总纠结“大数据”中的这个“大”字，如果非要

关注，也不要仅限于“量的大小”，而是理解成“大的变革”。也就是说，传统学科所面临的“数据”有了“大的变革”。随之，各学科要做的工作、要用的方法及要面对的问题也需要变更。可以这样理解，大数据时代到来之前，每个学科对数据都有自己固有的一套认识和处理方法。但是，大数据时代的到来，迫使人们改变这些传统认识。

数据变了，与每个学科中固有的数据认识论不同了。原来人们一直以为数据是“那样”的，但现在却变成“这样”了。以社会科学为例，以前人们都是挑选一些关键节点进行数据采集，如小区进出口有保安，登记你的姓名，进去了就没有其他记录了。现在，小区门口和小区内都有摄像头，采集的数据比较全面，那么，这种数据又如何处理和分析呢？在传统理论中找不到答案。这就需要一个新的理论——大数据理论，即数据科学。

大数据不是小数据的“简单集合”。从“小数据”到“大数据”的过程中产生了“涌现”现象。“涌现”才是大数据的本质特征。所谓“涌现”，就是系统大于元素之和，或者说系统在跨越层次时，发生了新的属性或新的质变。例如，大数据中个别数据可以有误，允许缺失、冗余、垃圾数据的存在，但不影响大数据的质量；再如，大数据中的每一条数据都“没什么用”，但放在一起就“很有用”；大数据中的每一条信息都“不是什么秘密”，放在一起“就得保密”。也就是说，从小数据到大数据，会涌现出很多你想象不到的特征。

回到“大数据”这个话题，以大数据为例说明“涌现”现象。比如，交通大数据，街上有很多摄像头，交通部门收集了大数据，你如果向交通部门要数据，他们说保密，不能提供。你可能会很郁闷，大街上发生的事情是公开的，摄像头也是公开的，摄像过程也是公开的，怎么到他们那里就“保密”了？其实，从数据科学的角度讲，交通部门的做法是合理的。交通部门的每一条数据都不是什么秘密，但是这些不保密的数据放在一起，就不得了了，从中可以分析出你的行为习惯，涉及个人隐私、社会安全甚至国家安全。这就是大数据的“涌现”，也是用基于“小数据”的理论不能解释“大数据现象”的原因所在。

新闻学与大数据交叉后，产生了一门新的研究领域——Data Journalism。金融和大数据交叉之后，出现了“大数据金融”，很多学科中都出现了一个新的方向。那么，这些新的学科交叉会出现什么？或者说，这些新的学科中有

哪些共同性的理论呢？那就是数据科学。也就是说，数据科学将会是学习这些领域知识的基础理论。

图灵奖的获得者 JimGray 提出的科学研究的第四范式又称《数据密集型科学发现》(*Data-intensive Scientific Discovery*)。在他看来，人类科学研究活动经历过三种不同范式的演变过程（原始社会的“实验科学范式”、以模型和归纳为特征的“理论科学范式”和以模拟仿真为特征的“计算科学范式”），目前正在从“计算科学范式”转向“数据密集型科学发现范式”。第四范式（“数据密集型科学发现范式”）的主要特点是科学研究人员只需要从大数据中查找和挖掘所需要的信息和知识，无须直接面对所研究的物理对象。例如，近年来天文学家的研究方式发生了新的变化——从海量数据库中发现所需的天体活动的照片，而不再需要亲自进行太空拍照。那么，JimGray 提出的第四范式对我们的科学研究有什么意义？在于绝大部分大学生的研究范式有待调整——他们往往习惯性地“采用问卷调查等方法亲自收集新数据”，而不是“首先想到有没有现成的大数据及如何再利用已有的数据（数据洞见）”。在大数据时代，研究范式需要调整，需要学习的专业理论、方法、技术、工具、最佳实践等都得拓展，甚至必须改变。这就是我为什么说“数据科学是现代人才的必修课”的原因所在。

作者简介

朝乐门：1979 年生，男，中国人民大学数据工程与知识工程教育部重点实验室、信息资源管理学院副教授，博士生导师；章鱼大数据首席数据科学家。中国计算机学会信息系统专委会委员、ACM 高级会员、国际知识管理协会正式委员、全国高校大数据教育联盟大数据教材专家指导委员会委员；主持完成国家自然科学基金、国家社会科学基金等重要科学研究项目 10 余项；参与完成核高基、973、863、国家自然科学基金重点项目、国家社会科学基金重大项目等国家重大科研项目 10 余项；获得北京市中青年骨干教师称号、国际知识管理与智力资本杰出成就奖、Emerald/EFMD 国际杰出博士论文奖、国家自然科学基金项目优秀项目、中国大数据学术创新奖、中国大数据创新百人榜单、全国高校大数据教育杰出贡献奖等多种奖励 30 余项。

朝乐门是我国第一部系统阐述数据科学理念、理论、方法、技术和工具的重要专著——《数据科学》(清华大学出版社, 2016)的作者,也是数据科学与大数据技术专业第一个领域本体“**Data Science Ontology**”的研发团队的总负责人。2017年9月,出版《数据科学理论与实践》(清华大学出版社, 2017)。

数据科学与大数据技术——新工科、新探索

上海工程技术大学电子电气工程学院院长 方志军

目前，数据科学与大数据技术专业的申报、建设，正好与我国新工科建设同步展开。数据科学与大数据技术专业是典型的新工科专业，在新工科背景下，如何建设好数据科学与大数据专业，要以人才培养需求为导向不断进行探索。2017年7月13日，上海工程技术大学电子电气工程学院院长、教授方志军在CIO时代APP微讲座栏目作了题为《数据科学与大数据技术——新工科、新探索》的微讲座。

一、新工科建设的进展

2017年2月18日，教育部在复旦大学召开了高等工程教育发展战略研讨会，共同研讨新工科的内涵特征、新工科建设与发展的路径选择，形成了“复旦共识”。2017年4月8日，教育部在天津大学召开了新工科建设研讨会，60余所高校共同商讨新工科建设愿景与行动，形成了新工科的建设行动路线，也就是“天大行动”。2017年6月9日，教育部在北京召开了新工科研究与实践专家组成员成立仪式暨第一次工作会议启动仪式，系统部署了新工科建设，对新工科的建设提出了指导意见，并发布了新工科研究与实践项目的指南。以“复旦共识”“天大行动”“北京指南”为新工科建设的三个重要时间节点，新工科研究的课题申报工作在全国范围内陆续展开，数据科学与大数据技术专业建设也处于不断探索的过程中。

二、新工科建设特色

（一）办学特色

随着产业的发展，上海工程技术大学在办学过程中成立了很多行业或产

业学院，如汽车学院、航空学院、轨道学院、服装学院等。电子电气工程学院是一个学科性学院，有电子电气、自动化、控制、计算机、大数据等专业，学院办学完全继承了上海交通大学的办学传统，学院的学科和人才培养与行业企业紧密相连，服务了其他行业特色的需要。

（二）专业特色

在数据科学与大数据技术专业建设的过程中，电子电气工程学院进行了以下探索：①树立“新工科”建设理念。大数据的专业建设要主动适应新经济的发展，探索教育教学的新思路。②设置“Computer+X”新结构。大数据专业设置在计算机系，有别于有的学校设置在数学系、统计系或其他创新班，是在计算机核心知识结构与框架的基础上展开的。③探索“科学—技术—工程”一体化的协同创新体系。不仅瞄准行业背景，在进行工程实践的同时，对大数据前沿问题进行探索和研究。④打造“世界一流大学+上海工程技术大学+行业知名企业”多元化的协同办学模式。依托上海国际化大都市的优势。与行业企业紧密结合，同时引进世界一流大学的教授团队，形成多元化的教学科研团队。⑤构建“课内课外+校内校外+镜内镜外”多维化的协同育人平台。这些新理念、新结构、新体系、新模式和新平台也为学院和学校的新工科建设奠定了良好的基础。

（三）课程体系

课程体系方面，首先，电子电气工程学院构建了全学院学科基础平台课程，设置有软件、数据结构、概率、统计等与大数据紧密联系的课程；其次，通过开设电路、电子等课程，在硬件知识体系构建过程中形成了自身的办学特色；另外，数据科学与大数据技术专业设置在计算机系，操作系统、数据库等核心课程得到体现；此外，实训课程的设置与行业企业良好结合，如交通大数据、轨道大数据、服装大数据、视频大数据等。

根据人才培养目标，结合学校与学院的学科特色，数据科学与大数据技术专业的培养计划由公共基础教育、学科基础教育、专业教育、集中实践教学四个模块组成，具体如下。

1. 公共基础教育（59.5 学分）

公共基础教育按照工程人才培养的共性要求和培养高素质人才的要求而设置，并为推进全面素质教育奠定基础，包括综合基础和综合选修两个模块。在综合基础模块中设置了由思想政治类、军事类、体育类、高等数学类、物理类、外语类等必修课程，在综合选修模块中设置了人文科学与艺术、社会科学、自然科学、创新实践等构成的选修课程。重点培养学生的数学应用能力、良好的沟通能力、表达与写作能力及基本工程与科研素养。

2. 学科基础教育（36 学分）

学科基础教育培养大数据人才必备的学科基础理论知识与实践能力，主要包括工程数学类、计算机类、电子技术类课程。工程数学类课程主要培养学生的线性代数与统计学基础；计算机类课程开设必备的计算机基础课程及课程设计，培养学生坚实的算法基础与程序设计能力。具体如表 1 所示。

表 1 基础课程及学分设计

课程名称	学 分	总 学 时	建议修读学期	学分要求
高级语言程序设计 A	3	48	1	19
数字电子技术	3	48	2	
线性代数	2	32	2 下	
电路（一）	2	32	3 上	
电路（二）	2	32	3 下	
电路实验	1	20	3 下	
概率论与数理统计	3	48	3	
微机原理及接口技术	3	48	4	
离散数学	3	48	3	7
算法与数据结构	4	64	4	
面向对象程序设计	4	64	4	4
数据科学与大数据分析导论	2	32	3	6
Linux 操作系统	2	32	4 上	
数字图像处理	2	32	4 下	

3. 专业教育（31 学分）

分为专业必修课程和专业选修课程两个模块。通过专业必修课程的学习，

掌握数据科学与大数据技术的核心知识，并具备一定的数据获取、挖掘与分析的能力；通过专业选修课程的学习，进一步拓展处理实际问题与可视化展示的能力。具体如表 2 所示。

表 2 专业课程及学分设计

课 程 组	课程名称	学 分	总 学 时	建议修读学期	学分要求
必修课程	(分布式) 操作系统	4	64	5	25
	(大型) 数据库原理	4	64	5	
	高级应用统计学	4	64	6	
	数据挖掘与分析	4	64	6	
	数据采集与搜索技术	3	48	5 上	
	软件工程	2	32	7 上	
	人工智能	2	32	6 下	
	高级算法分析	2	32	6 下	
专业选修课	Java 程序设计	2	32	5 上	6
	R 语言编程	2	32	5 下	
	MATLAB 编程	2	32	5 下	
	并行程序设计	2	32	5 下	
	统计软件	2	32	6 上	
	数据可视化分析	2	32	6 下	
	商业智能	2	32	7 上	
	计算智能	2	32	7 上	
	智能推荐技术	2	32	7 上	
	机器学习	2	32	7 上	
	计算机视觉	2	32	7 上	
	(Hadoop) 大数据存储与运算	2	32	6 下	
	云计算与虚拟化技术	2	32	6 下	

4. 集中实践教育（31.5 学分）

为了提升学生的工程素养，强化培养学生的工程能力，构建了由校内学习实践、企业学习实践和创新与综合工程素质培养三部分构成的实践能力培养体系。表 3 所示为实践课程及学分设计。校内学习实践和企业学习实践共同构成集中实践教育环节，在培养计划中集中统一安排。创新与综合工程素质培养教学环节安排在课外，由学生根据本人的兴趣爱好与特长自主选择参加。

表 3 实践课程及学分设计

课程名称	学 分	总 学 时	建议修读学期	学分要求
军训	(1)	(2) 周	1 上	31.5
认知实习	1	1 周	1 下	
合作教育（一）	(2)	(6) 周	2 下	
制造技术基础实习	2	2 周	2 下	
高级语言程序设计课程设计	1.5	1.5 周	2 下	
合作教育（二）	2	(6) 周	4 下	
电工实习	2	2 周	4 上	
数据采集与搜索综合实验	2	2 周	5	
大数据分析可视化综合实验	2	2 周	6	
大数据系统应用综合实验	2	2 周	7	
合作教育（三）	2	(6) 周	6 下	
毕业设计（论文）	16	16 周	8	
视频大数据综合实验	2	2 周	7 下	2
智能交通大数据综合实验	2	2 周	7 下	
社交媒体大数据综合实验	2	2 周	7 下	
金融大数据综合实验	2	2 周	7 下	
教育大数据综合实验	2	2 周	7 下	

集中实践环节中，通过数据采集与搜索综合实验、大数据分析可视化综合实验、大数据系统应用综合实验，涵盖了整个大数据处理流程。同时在第七学期安排多个面向应用的综合实验，与产学合作基地联合培养学生，着力提高学生的工程意识、工程素质、工程实践能力、创新精神、职业技能、职业道德、团队协作精神和交流能力。

在人才培养过程中，学院注重与世界一流大学及行业企业的合作与交流。在具体教学过程中，倡导前沿知识进课堂、科研成果进课堂、企业实践进课堂、课程思政进课堂、人文社科进课堂，实现全方位、全过程育人。

三、关于上海工程技术大学

上海工程技术大学是我国新工科建设地方高校组的牵头单位。之所以会

成为地方高校组的牵头单位，与学校的办学历程和办学特色息息相关。上海工程技术大学的办学历史，可以追溯到 1978 年成立的上海交通大学机电分校，学校从办学之初便是工科教育。电子电气工程学院的目标是全面瞄准工程技术前沿，主动对接国家战略需求，积极服务区域经济社会发展，集聚高端人才，彰显培养特色，逐步把学院建设成同类高校领先的高水平电子电气工程学院。

作者简介

方志军：上海工程技术大学电子电气工程学院院长、教授、博导。IEEE/ACM 高级会员，中国计算机学会多媒体技术专业委员会委员，中国人工智能学会智能服务专业委员会委员。先后承担了国家自然科学基金项目、公安部应用创新计划项目、江西省对外科技合作项目、上海市科委地方能力建设等项目等多项课题的研究，担任了多个学术会议的组织委员会主席或共同主席。

当“数据科学”遇上“自由博雅”

同济大学计算机科学与技术系副教授 王 伟

一、大数据时代的冲击

大数据时代给人们的工作、学习和生活带来了全方位的冲击。

（一）思维模式

大数据作为继云计算、物联网之后 IT 行业又一颠覆性的技术，备受关注已是毋庸置疑的事实。大数据就像 21 世纪的石油和金矿，是一个国家提升综合竞争力的又一关键资源。大数据既是一类数据，也是一项技术，还是一种理念。关于大数据的理念、大数据的原理、大数据的应用，每个人都应该或多或少地了解、掌握。特别是大数据的思维方式，个人认为，大数据的理念和思维方式已经成为人们应该具备的基本常识。

（二）课程教育

从 2015 年国务院常务会议通过的《关于促进大数据发展的行动纲要》非常强调开发应用好大数据这一基础性战略资源。教育部高等学校教学指导委员会也将“大数据”列为“十三五”期间高等学校的教学改革和教学建设的重点，同时教育部高等教育司也于 2016 年在普通高等学校本科专业设置中增加了《数据科学与大数据技术》专业（专业代码为 080910T），以及面向高职院校的《大数据技术与应用》专业（专业代码为 610215）。新专业的设置为目前国内高校的学科布局带来了新的挑战与机遇。

（三）技术体系

自从大数据出现以后，数据管理界发生了巨大的变化，技术驱动成为大数据管理系统的一个主要变革力量，并开始走向成熟。大数据技术的软件栈

也基本成型，得到了业界与学术界的基本认可。大数据目前的几个主要的发展趋势包括：分布式计算已逐渐成为主流计算方式；数据分析算法逐渐丰富的同时，工具逐渐普及化，Hadoop、Spark 及其生态系统将重构数据处理市场，以及大数据产业链日益繁荣等。

因此，“数据强国”已经上升到了国家的战略高度，国家领导人也在不同场合多次指出：

“当今世界，科技进步日新月异，互联网、云计算、大数据等现代信息技术深刻改变着人类的思维、生产、生活、学习方式，深刻展示了世界发展的前景。”

“因应信息技术的发展，推动教育变革和创新，构建网络化、数字化、个性化、终身化的教育体系，建设‘人人皆学、处处能学、时时可学’的学习型社会，培养大批创新人才，是人类共同面临的重大课题。”

二、数据科学：从大数据到行动

作为大数据背后的学科基础，数据科学可以被看作“思维+计算机科学+统计+应用”的一个综合体。它包含三个层面的意思：

- 首先，建立数据思维方式，学习怎样利用数据。
- 其次，应该了解数据清理、集成、探索等相关技术。
- 最后，洞见和商业意识也至关重要。

（一）数据科学的三大支柱

数据科学天生就是一个交叉学科，与数据科学最为密切的学科有计算机科学与技术、数学、统计学、信息管理、情报学等。因此，数据科学的三大支柱可以归纳如下。

- Datalogy（数据学）：对应数据管理（Data Management）。
- Analytics（分析学）：对应统计方法（Statistical Method）。
- Algorithmics（算法学）：对应算法方法（Algorithmic Method）。

（二）数据科学的五大要素

在此基础之上，我们凝练了数据科学的五大要素，并用一个名为 A-SATA 的模型来表示，包括分析思维（Analytical Thinking）、统计模型（Statistical

Model)、算法计算 (Algorithmic Computing)、数据技术 (Data Technology) 及综合应用 (Application)。这些是数据科学中关键的内容。

(三) 数据科学的核心知识点

从这个模型可以导出数据科学的核心知识点,这也是后来课程建设的关键。

- 分析思维 (Analytical Thinking): 包括计算思维 (Computational Thinking) 和统计思维 (Statistical Thinking)。
- 数学基础: 微积分、线性代数、概率统计、离散数学等。
- 数据建模与评估: 统计模型、回归模型、模型评估等。
- 算法实现: 问题求解能力和算法等。
- 数据管理: 涉及数据的整个生命周期,包括感知、存储、计算、分析、可视化等。
- 知识转化: 沟通交流、道德规范等。

实际上,目前国内外很多大数据相关学院、专业都是围绕上述核心知识点来进行课程开发和整合的。同济大学也是如此,例如,这学期开设的《数据科学通识导论》课程便是围绕上述内容展开的。

三、对于数据科学的思考

对大数据和数据科学另外一个重要的思考是希望能将它们作为通识课来进行教授。

首先,数据科学(包括大数据)非常适合作为一门通识课程。所谓通识课程,旨在为学生带来完整的知识结构,养成触类旁通的通用智慧。通识教育起源于古希腊,指公民应当具备的知识与能力。比起“知识点”,通识课更侧重于“思考点”和“实践点”,因此,它强调的是以学为中心,老师会想方设法激发学生自主学习的动机,让学生自觉地收集资料、思考问题、表达观点、自我辩护、开展实践,在实践中使思考能力、语言能力及动手能力合一。这些训练都是培养视野广阔、人格完整、智识践行的人才不可缺少的。这正是通识教育的意义。

基于这种思想,大数据和数据科学实际上是非常适合作为一门通识实践课程的。

- 数据科学有利于培养信息时代的人才——因为在信息时代，和数据信息打交道的场景无所不在。
- 数据科学有利于培养跨学科视野——数据科学的本质便是跨学科。
- 数据科学有利于培养表达自我所必备的技能——目前的信息时代，包括图表技能都是一种很好的表达能力。
- 数据科学有利于培养个人的科学思维方式；特别是数据科学中的量化思维和计算思维。无论是理工科还是人文社科类的学生都应该掌握。

数据科学有利于围绕数据开展实践。因为数据已极大丰富，获取数据已成为越来越容易的事情。

“通识教育”中有一类特殊的课程是每个学生都必修的，这就是“自由教育”的课程，其精神支柱和思想来源就是现在经常听到的“Liberal Arts”，我们将这个词翻译为“自由博雅”。“自由博雅”通常包括自然科学（Sciences）、社会科学（Social Sciences）和人文学科（Humanities）三部分。Liberal Arts旨在培养一流的头脑、一流的心灵。只有有了一流的头脑与心灵，才有可能产生一流的科学家、艺术家和思想家。否则，大学生产的只是一群高级工匠，知其然而不知其所以然。相对于具体的职业教育而言，Liberal Arts的目标不在于教会学生某些具体的谋生的技能，而是从多方面对学生进行教育，使其成为一个高素质、有教养的文化人。上述几个方面都和数据科学的“世界观”不谋而合，Liberal Arts理念可以向数据科学注入“博雅”之心，为数据科学带来：

- 问正确问题的能力。
- 科学方法观。
- 团队协作的精神。
- 沟通交流的能力。
- 三观正确的决策。

而这些，也都是数据科学所应该关注的。

四、数据科学课程建设与教学实践

因为本人在高校的原因，下面着重介绍数据科学在课程建设方面的内容，并结合自己的教学实践谈谈心得体会。从2016年开始，笔者在学校陆续开设

了两门与数据科学（包括大数据）相关的导论类课程：“大数据原理与实践”和“数据科学通识导论”。

（一）课程内容的设置

以“数据科学通识导论”为例，我们围绕前面提出的 A-SATA 的模型构建了这门课程的知识体系，包括思维概念、数据技术、数据分析、算法编程和综合应用五大模块，共 16 节课，基本覆盖了上述数据科学的核心知识点，如图 1 所示。

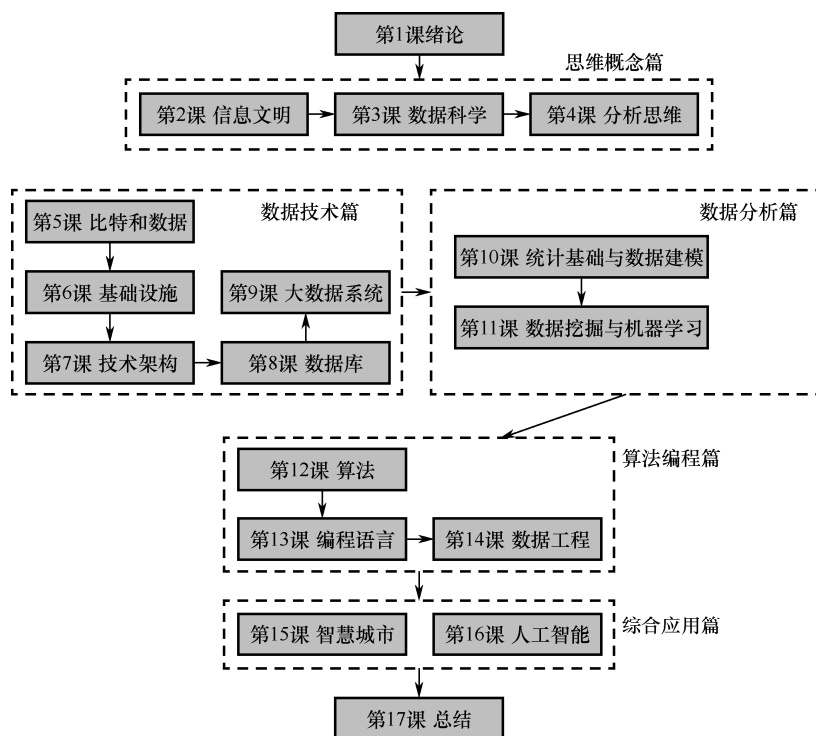


图 1 “数据科学通识导论”课程知识体系

（二）“自由博雅”实践

重点开展了“自由博雅”的实践，总结起来共 16 个字：建立对话；激发思辨；协作交流；动手实践。这里，借助了两个平台：微信公众号平台（微信号：Datahui）和数据科学实践平台。

1. 微信公开课实践

为了更好地服务学生，以及体现通识导论课程的性质，我们借助了“微信”这个强大而普世的平台，开展师生之间的连接和互动。例如，每周课程的更新方式如下。

- 周二：发布本周课件初稿，授课。
- 周三、周四：互动、点评与问答。
- 周五：发布最终版课件及相应的文本注释。
- 周末：课外阅读文章。

希望通过这种方式，激发出学生与老师的潜力，并很好地连接老师和学生，围绕数据科学进行师生互动，结合课程思考题，达到建立对话、激发思辨的目的。

2. 数据科学实践平台

实践出真知，数据科学者们所教授的课程尤其如此。因此，我们采取了多种方式，尽量给学生提供数据及动手的机会，达到协作交流、动手实践的目的。这里主要包括如下三类。

- 课内的课程设计。
- 引入课外竞赛：Kaggle、上海 SODA、阿里巴巴天池等。
- 数据马拉松（Datathon）：类似于 Hackthon，以集中的时间完成项目。

结果还是比较令人满意的，学生做出了一些非常令人赞叹的作品，例如，上海地铁系统进站流量图、基于人流指数预测的商圈公共安全预警系统、轨道交通运维大数据分析等。部分作品获得了一些相关比赛的奖励，也产生了教研结合的后续项目。

五、大数据实践平台建设

数据科学的实践需要一个非常好的平台，才能为老师和学生提供实践服务。着眼大数据本身，希望能引入校企合作，我们目前正与上海的大数据高科技公司共同建设大数据实践平台。当前的大数据还是技术驱动型的，很多技术还不完善，高校研发力量无法跟上国家大数据的发展，需借助一些企业

的力量推进相关研究（见图 2）。

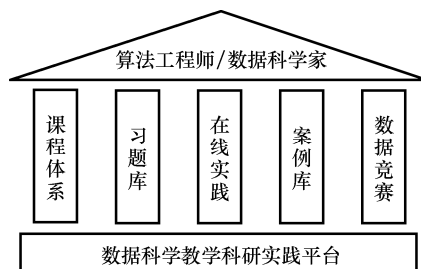


图 2 数据科学教学科研实践平台

数据科学教学科研实践平台包括课程体系、习题库、在线实践、案例库及数据竞赛，这些还是比较容易解决的，但大数据还需强大的基础设施来指导实践，随着大数据的成熟，在平台建设方面会更加便捷。

六、总结：未来的挑战

目前高校的数据科学实践平台还非常依赖于技术，愿景是好的，挑战也颇大，特别是在技术上：数据上云、分析上云、教育上云等，同时，容器技术、HCI 技术、分布式处理技术等也是很大的挑战。

愿景是希望未来建设 P 级的大数据公共实践教学平台，达到一个学校甚至是一个区域资源优势共享的目的。

作者简介

王伟：同济大学计算机科学与技术系副教授，博士生导师，CCF 高级会员，CCF 教育工作委员会委员、CCF 大数据专委会通信委员、CCF 体系结构专委会委员。美国 Wisconsin 大学 Madison 分校高级访问学者，美国 Florida 大学 CSC 访问学者。研究方向为云计算、大数据处理与大规模在线学习系统。

数据科学与大数据技术专业申报与建设

中南大学信息院教授 张祖平

一、市场需求

从目前来看，大数据毋庸置疑应该作为一个专业来申报，但最初还是有很多声音存在的。在申报时，一般要从国际和国内市场需求的角度及发展形势进行分析，网上有很多材料可供参考。这里重点介绍区域性的作用和要求，例如，中南大学坐落在湖南省长沙市，如果大数据专业申报成功的话，需要思考这将为湖南省带来什么样的影响，为长沙市带来什么样的影响，如此相应地也就会响应区域性的政策与相关的发展导向。

就整个社会对人才的需求而言，现在网上有很多统计数据在响应这个话题，如现在急缺多少大数据人才、有多少个岗位需要这种人才等，但这仅仅是专业申报时的一种说法。从考生自身或家长的角度来看，毕业时能否顺利找到好的工作才是他们最关注的。2015 年，中南大学计划在计算机科学与技术专业中建立大数据方向，计划招收两个班，一个班来自计算机学科，另一个班来自其他相关工科专业。但实际情况是，2015 年大数据还是相对比较冷门的，很多人还对这个专业的认识不够，不确定其市场的接收程度，因此，最后只招到了 35 个人。在 2015 年申报完成数据科学与大数据专业后，2016 年的正式招生情况有明显的好转，面向全国招生的数据科学与大数据技术专业的分数明显高于计算机及信息安全专业，这表明大家已开始认同这个专业。2016 年全国有 3 所学校获得了大数据本科专业的招生权，到 2017 年增加了 32 所学校，共有 35 所高校有数据科学与大数据技术专业可以招生，专业已经得到政府部门的认同。2017 年申报大数据本科专业的院校有 255 个之多。

二、指导思想

原来在申报时，国内一直在讨论这个专业是理科专业还是工科专业，或者是理工相结合的专业，涉及专业的培养定位问题。由于数据科学与大数据技术考虑的是理论与技术结合，应该既有理论又有技术，属于理工相结合的专业。

数据科学与大数据技术专业的定位是以现代计算机与网络系统为依托，专注大数据采集与管理、大数据分析与应用的理论和技术，培养解决大数据系统建设整体性高级复合型人才，同时能够承担企事业单位、政府部门、社会组织的信息分析与管理系统、信息咨询服务平台、信息共享网络等项目的专业技术工作。当时规划第一年招收 60~70 人，5 年内扩展到 90 人。其中需要考虑对已有专业与传统专业的格局影响问题，这个专业既与计算机专业相关，又与软件工程相关，因此，需要与这两个专业形成优势互补的关系。另外，随着大数据专业的兴起，还有一些传统的老专业，如医学信息学、信息管理与信息系统、审计学、情报学、管理信息系统等，慢慢会被融合进来，最终被替换掉。

三、培养方案

培养方案当中也涉及定位问题，一方面要结合新的发展趋势，另一方面要考虑自身的条件。具体包括以下几个方面。

（一）知识要求

大数据需要具有较好的数学基础，数学方面要有核心的课程，同时还应具有一些具体领域的基础，这就需要一些相关特色领域的业务内容。当然，本身的计算机基础也是需要的。因此，我们规划了一些课程，如计算思维和数据科学等与数学相关的课程。与具体领域相关的如大数据与领域建模、医药信息系统等。与大数据直接相关的如数据采集技术、云计算与数据中心、医疗大数据等。计算机方面如一些大型数据库技术、数据可视化技术、机器学习与模式识别、非结构化数据处理技术、分布式海量存储系统、数据安全等。

（二）能力要求

（1）要具备大数据应用系统的设计与实现能力，特别是在数据分析、数据管理、数据存储等方面，应该受到较为系统的工程训练，能发现、分析和解决实际工程技术问题。

（2）具备良好的工程项目交流、表达、组织、管理、协调与沟通能力。例如，在做实验的过程中，原来强调个人独立训练，目前在课程设计和环节中强调团队配合。

（3）了解信息科学、计算机学科、数据科学的发展动态，并掌握相关文献检索方法，具有基本的专业资料分析与综合能力、良好的文档与科学论文撰写能力。

（4）具有较强的创新意识和一定的创新创业能力。

（三）素质要求

素质要求包括道德修养、集体主义精神、理想信念等。

在培养方案中，起初规划是 180 学分，其中必修是 136.5 学分，选修是 43.5 学分，具体体现为几个模块，如公共基础课程、学科基础课程、专业课程、素质拓展环节等。

主干课程，如信息论与编码、计算思维和数据科学、离散数学、数据结构、操作系统、数据库原理、计算机网络原理、数据挖掘、数据安全等。特色课程，如数据采集技术、云计算与数据中心、机器学习与模式识别、大型数据技术、数据可视化技术、图像视频与非结构化数据、分布式海量存储系统、大数据与领域建模、医药信息管理、医院信息化、医疗大数据等。最后在正式实施时，由于学校统一的学分要求，最终确定为 160 学分。

四、学科基础

这里涉及中南大学的一些特色实验室，如 2012 年的教育部“移动医疗”重点实验室、2013 年的湖南省金融货币识别与自主服务平台工程技术研究中心、2010 年的湖南省区域医疗信息共享与协同服务示范平台、2014 年的声探

测与信息对抗湖南省国防科技重点实验室、2015 年的湖南省“医学大数据”协同创新中心、2016 年的网络资源管理与可信评估服务湖南省重点实验室、2017 年的医疗大数据应用国家工程实验室（共建）。

学科是专业的支撑，中南大学相关的计算机学科名列前 20 位（第三轮，最新的第四轮是前 10 名），ESI 进入全球 1%。另外，还有一些国家、省级奖励作为支撑。学科基础对专业支撑很重要，由于有相关的特色学科及相关平台支撑，所以第一批申报能获得成功。

五、师资队伍

这是一个看起来没问题但实际有问题的方面。当时在申报专业时，软件学院、信息安全与大数据研究院等均有参与。目前在中南大学，相关的教师有 100 多人，其中信息科学与工程学院的计算机科学与技术系及软件学院有 90 多人，信息安全与大数据研究院有 80 多人，网络信息中心有 30 多人。从办学结构来看，有专业的教师、博士、学者等作为支撑。但在建设过程中，还存在复杂的问题，主要是将新的课程安排下去不是一件容易的事情。

六、基本条件

首先是实验室，如果原来是计算机专业来支撑大数据专业，那还好，有很多台式计算机及服务器，还有云计算与大数据平台作为服务，还需要科研的相关平台来支撑。大数据专业的实验室是烧钱的，因为大家都知道投入比较大，投入小做不成像样的大数据实验室。我们申报的时候算有一些基本条件，当然还包括一些教学内容的筹备，主干课程、特色课程等教学大纲的处理等工作。

七、办学特色

第一，具有医学与医药、轨道交通、有色金属工业领域的行业特色和优势。

第二，具有“湖南省医学大数据协同创新中心”“医疗大数据平台”“轨道交通大数据平台”等基础设施方面的实践性教学特色与优势。

第三，专业本身面向社会对人才的需求，创立了人才培养体系。

第四，具有交叉型学科群与人才团队合作的优势。多学科多学院合办大数据专业，与已有专业交叉融合，促进传统专业发展。

第五，特色系列教材建设与特色学科方向。具体有透明计算与主动服务、计算优化及应用、计算机视觉与数字医疗、可信计算与计算机网络、网构软件与网络资源管理、数据科学与医学大数据，主要是医疗大数据全面集成与融合，形成了真正的医疗大数据环境。

下面介绍两个相关的平台。

医疗大数据国家工程实验室包括医疗大数据标准、智慧终端、大数据系统、网络搜索、智慧管家、智慧医院等核心技术。湖南省“医学大数据”协同创新中心主要做了“1 中心 4 平台”：医学数据科学理论与技术研究中心、医学大数据集成共享平台、医学大数据处理分析平台、医学大数据应用研究与创新平台及大数据驱动网络信息服务平台。如果没有相关平台作支撑，创办这个大数据专业是很困难的，一方面我们充分利用现有互联网上的数据，另一方面要从科研平台中挖掘相关的数据。

网络资源管理可行性评估重点实验室，主要针对目前的互联网、物联网中的数据感知、数据关联及资源的监管、交易、封装、评估和发布等，面向科技大数据、智慧城市、智慧工厂、智能制造等。将来办学进行的大数据分析，需要有一个落脚点，包括面向全国共享的云平台，另外，也需要自身的学科支撑，具备大数据来源和分析平台才能站住脚。

八、课程调整

一开始我们总共做了 34 门课程，包括实习、实训、语言的课程，随着专业的批复和正式招生，落实到老师时产生了很大的困难。从现实的角度，目前重点建设 16 门课，包括新生课、数据科学与大数据技术导论、数据采集与融合技术、信息组织理论与技术、科学计算与数据建模、数据仓库与数据挖掘、Python 数据处理编程、R 语言数据分析编程、信息组织课程设计、分布式系统与云计算、数据处理方法课程设计、智能搜索引擎技术、医疗大数据、大数据编程、大数据综合应用实践及深度学习。其他的课程在计算机科学与

技术专业和大数据专业开设，部分课程因种种困难最后拿掉了。

九、专业实验中心建设

实验室的投入是很大的，但可以利用科研平台的共享来支撑大数据专业的教学与实验。我们首期投入 90 万元，已经完成了招标，正在落实建设，具体在大数据计算资源池、大数据实践教学管理平台专用服务器、VDI 并发授权、大数据实践教学管理平台、大数据可视化分析教学资源、R 语言教学资源、大数据实践教学系统环境等方面展开，长远规划了大数据专业实验中心，投入近 600 万元，各个学校可以根据自身条件选择性地建设特色实验室。

作者简介

张祖平：基础数学本科、硕士，计算机应用专业博士，中南大学教授，博士生导师，基础医学博士后，加拿大西安大略大学(The University of Western Ontario)国家公派访问学者。计算机科学与技术系主任，大数据与知识工程研究所副所长。中南大学“531 人才计划”第二层次人才。长期从事大数据与知识工程、信息度量与信息融合、软件工程与信息系统、参数计算与生物计算等方向的研究。项目查重与文献甄别系统推广应用到 10 多个国家级及省部级单位。发表学术论文 100 多篇，主编《信息学科导论》《数据库原理与应用》《计算机网络技术教程》等教材。主持国家自然科学基金项目面上项目 2 项，委主任基金 2 项，企事业单位大型软件工程或信息化项目 10 多项，多项大型软件系统通过国家或省级鉴定，获省部级科技进步一等奖 2 项、二等奖 1 项、三等奖 2 项。

关于高校大数据教学若干关键问题的探讨

厦门大学信息科学与技术学院系助理教授 林子雨

一、如何搭建大数据实验平台

在高校大数据教学过程中，实践是很重要的环节，因此，很多高校在开设大数据课程时，不仅要选择好的教材，同时也要选择好的大数据实验平台。总体而言，当前国内有以下几种大数据实验平台方案。

（一）建设统一的大数据实验机房

目前，在国内有如下两种典型的大数据实验机房建设方案。

第一，多台终端机采用云桌面方式连接到中心服务器。这种模式在一部分高校已得到广泛使用。通常而言，中心服务器采用高密度服务器，采用虚拟化技术得到很多虚拟化资源，所有的终端机都可以连接到中心服务器，共享这些虚拟化资源，终端机仅起到云桌面的作用，数据处理运行都是在中心服务器上进行的，因此，这种模式对终端机的配置要求较低。这种模式在实际的高校部署中，又会有两种不同的方案：一种方案是在学校本地机房放置中心服务器，各个终端机直接连接到本地服务器。另一种方案是中心服务器不是放在高校实验室的内部机房中，而是放置在阿里云等公有云平台上，或者放在大数据实验平台供应商的数据中心，高校可以通过浏览器访问云端的大数据实验环境。

第二，用多台物理机器构建分布式环境。在这种模式中，每台物理机器都构成一个分布式计算节点，多个节点构成分布式的集群环境。在这种模式下，高校的通常做法是将学生进行分组，如 5 个学生一组，为 5 个学生分配 5 台物理机器，由学生完成大数据集群环境搭建，或者机房管理员已经为这 5 台机器统一安装了大数据集群环境，学生可以直接在这 5 台物理机上进行相关的大数据实验。

（二）单机构建实验环境

既然已经有了统一的大数据实验机房，为什么还需要单机方式呢？主要有 3 个方面的原因。

（1）有些学校没有建设统一的大数据实验机房，需要教师和学生自己在计算机上安装大数据实验平台。

（2）学校有统一的机房，但是上机时间有限，学生需要在宿舍或者实验室进行大量的课后上机实践，需要在自己的计算机上安装大数据实验平台，随时实践。

（3）学校的统一机房本身就采用每台机器独立安装的方式。很多机房都具有统一的管理平台，可以把大数据实验平台制作成镜像，然后自动快速把镜像部署到机房的每台计算机上。

如果采用这种单机构建模式，通常对整个实验室内部单机配置要求较高。一般而言，学生或老师的机器大多数为 Windows 系统，单机安装方式的具体方法如下：在 Windows 系统基础上安装虚拟机软件，如 VMWare 或 VirtualBox，在虚拟机软件上安装 Linux 操作系统，再在 Linux 操作系统上安装 Hadoop 等大数据相关软件。这种架构对底层的硬件配置要求是比较高的，因为要同时运行 Windows 系统和 Linux 系统，此时对底层资源的消耗比较大，尤其是对内存要求较高。一般而言，如果采用虚拟机方式（不是双操作系统方式），则单机方式构建大数据实验平台时，单机配置至少需要 8GB 的内存，否则系统运行会很缓慢。单机环境如何快速部署到其他机器中呢？老师首先在自己的计算机上完成大数据实验环境的构建，再将其导出做成“镜像”，存放到云盘，供学生下载，学生在自己本地计算机的虚拟机软件中直接导入镜像，就可以生成大数据实验环境，直接使用，避免了烦琐的大数据实验环境搭建过程。

（三）实验室多机构建分布式环境

在实验室内部，完成一些相关科研数据的处理、分析，或进行一些大数据教学案例分析，学生或老师仅用自己的计算机构建虚拟机方式是无法高效处理分布式大数据实验的，因此，需要利用实验室内部 3~5 台机器搭建起真正的物理分布环境，使庞大的数据进行分布式物理计算，这也是学生或老师采用的多机分布式物理环境。

二、如何解决云计算与大数据课程的知识交叉

当前,很多高校都在开设大数据、云计算这两门课程,其中可能会遇到一些尴尬问题,如云计算和大数据两门课程知识点重合度高,没有合理地安排内容,尤其是 Hadoop 等大数据只是在两门课程中可能均有介绍,使得两门课程在大数据知识方面高度重合,于是两门课程的老师要花费大量时间多次讲解同一项技术。导致学生要重复学习同一种知识,也导致了老师相互之间的尴尬等问题。出现这个问题的原因是云计算教材的选择出现了问题。云计算课程在选择教材时,云计算教材中包含了大量的大数据知识,如 Hadoop 生态系统的各个组件(HDFS、HBase、MapReduce、Pig、Hive、Zookeeper 等)。而老师上课通常都是围绕教材进行的,教材写了什么内容,就讲什么内容。云计算教材中包含了大量关于 Hadoop 等大数据知识,任课教师就只能按照教材讲大量属于大数据的内容。那么,为什么会导致这一现象的出现呢?这就要从云计算和大数据的渊源说起。

(一) 云计算和大数据的渊源

云计算技术诞生于 2006 年,云计算最初主要包含两类含义:一类是以谷歌的分布式文件系统(GFS)和分布式并行编程模型 MapReduce 为代表的大规模分布式并行计算技术;另一类是以亚马逊的虚拟机和对象存储为代表的“按需租用”的商业模式。也就是说,通过网络以服务的方式为用户提供非常廉价的 IT 资源这样一种商业模式,就像今天的百度云盘、阿里云等,都属于这种商业模式。所以,早期的一些云计算教材,就会包含上述两类内容,因此,会包含大量介绍 Hadoop 等大数据技术的知识。但是,随着大数据概念的提出,云计算中的分布式计算技术开始更多地被列入大数据技术,所以,现在人们提到云计算时,更多指的是底层基础 IT 资源的整合优化,以及以服务的方式提供 IT 资源的商业模式(如 IaaS、PaaS、SaaS),而很少会去谈及 Hadoop 等已经被单列为大数据的技术。

正是因为上述原因,在 2010 年左右出版的一些云计算教材,通常都会包含虚拟化、数据中心、分布式存储 GFS 和分布式处理 MapReduce 等内容,这

类教材被称为“云计算大数据复合型教材”，也就是说，虽然这类复合型教材的名称是“云计算”，但是，包含了大量讲解 Hadoop 等大数据知识的章节，而不是简略介绍 Hadoop。因为 2010 年之前，大数据和云计算的技术都是混在一起的，Hadoop 等大数据技术之前都被称为云计算。

2010—2014 年，云计算已经大规模普及，但是，大数据还没有大规模普及，因此，很多高校都没有开设大数据课程，只开设了云计算课程。这样，在 2010—2014 年，使用复合型云计算教材上课，当然是不会遇到问题的。但是，到了 2015 年左右，越来越多的高校开始开设大数据课程，这些大数据课程讲解的是 Hadoop 等大数据技术。这时，对于那些选用复合型云计算教材的高校而言，问题马上暴露出来，这些高校突然发现，由于前期选择了复合型云计算教材，导致大数据和云计算两门课程的内容重合度很高，两门课程的老师相互之间也很尴尬，到底如何协调好彼此的上课内容，显得很棘手。

（二）如何协调云计算和大数据两门课程知识点

如何解决这个两门课程知识点过高重合的问题呢？较好的方法是，在现在的云计算课程中不能继续使用复合型云计算教材（因为其中包含过多属于大数据技术的内容），也就是说，2015 年以后，如果一个高校同时开设云计算和大数据课程，在云计算教材的选择方面，最好把复合型云计算教材更换成“单一型云计算教材”。所谓单一型云计算教材，是指在云计算教材中，不能把 Hadoop 等属于大数据课程的内容作为核心内容，只用一个章节简单介绍 Hadoop 等大数据技术即可，不能用好几个章节进行大量介绍。也就是说，在“单一型云计算教材”中，Hadoop 等大数据技术只是教材的“次要内容”，只是为了考虑到云计算和大数据的紧密关系和历史渊源，才加以介绍，在实际授课时，用两个课时进行简单讲解即可。采用“单一型云计算教材”以后，就彻底解决了云计算和大数据课程知识点高度重合的问题，云计算课程只有两个课时的大数据技术简单介绍，详细的大数据技术，需要学生在大数据课程上通过 32 个学时来学习。

采用“单一型云计算教材”以后，云计算和大数据两门课程的知识重点就有了明显的区分，不会重合。

（1）云计算的教学重点：云计算概念、云计算体系架构、数据中心、虚

拟化技术（平台虚拟化、资源虚拟化、虚拟机的动态迁移、云操作系统）、SOA 架构及开发技术、云数据中心设计与测试、云数据中心维护与管理、云安全架构、桌面云、PaaS 应用开发平台、开源的云计算管理平台 OpenStack、Docker 容器、大数据存储与管理（最多两个学时，不需要实验，只是知识介绍）。

（2）大数据的教学重点：系统论述大数据的基本概念、大数据处理架构 Hadoop、分布式文件系统（HDFS）、分布式数据库（HBase）、NoSQL 数据库、云数据库、分布式并行编程模型 MapReduce、大数据处理架构 Spark、流计算、图计算、数据可视化，以及大数据在互联网、生物医学和物流等领域的应用。

三、如何建设优质的大数据教学资源

优质的大数据教学资源，直接影响大数据课程的顺利开设和大数据教学水平的发展。由于当前大数据教学正处于推广期，大数据教学资源还比较稀缺。同时，大数据知识体系非常庞杂，包含了数据生命周期内的各种技术，而且大数据知识更新换代非常快，类似 Hadoop 等大数据技术，刚兴起几年，又有 Spark 等新兴技术的崛起，这进一步加剧了教师开课的难度。

为了缓解高校大数据教学资源稀缺的现状，全国高校教育界同人都在不断努力，建设资源。全国高校大数据教育联盟多次组织召开大数据教学研讨会，组织相关高校教师共同开发教学资源。

（一）建立高校大数据课程公共服务体系的重要性

建立高校大数据课程公共服务体系，可以解决以下几方面的问题：

- （1）提供丰富的教学资源。
- （2）降低大数据课程开课门槛。
- （3）提升学生学习效果。
- （4）加快高校大数据课程建设进程。
- （5）不断提升高校大数据教学水平。

（二）案例

厦门大学数据库实验室致力于打造中国高校大数据课程公共服务平台

(主页), 建设了到目前为止国内高校最完备的大数据课程公共服务体系, 已经成为全国高校大数据教学知名品牌。平台以开放、共享的方式提供免费教学资源, 缓解大数据教育资源稀缺的问题, 降低大数据的开课门槛。目前, 平台建设了 11 个 1 工程, 包括 1 本教材、1 个教师服务站、1 个学生服务站、1 个公益项目、1 堂巡讲公开课、1 个示范班级、1 门在线课程、1 个交流群、1 个保障团队、1 个培训基地、1 个实验平台。访问厦门大学数据库实验室网站, 即可免费访问平台上的所有教学资源。

平台向全国高校免费提供开设大数据课程所需七大黄金资源。

1. 《大数据技术原理与应用》教材

笔者编著了国内高校第一本系统性介绍大数据知识的专业教材《大数据技术原理与应用》。教材系统论述了大数据的基本概念、大数据处理架构 Hadoop、分布式文件系统 HDFS、分布式数据库 HBase、NoSQL 数据库、云数据库、分布式并行编程模型 MapReduce、大数据处理架构 Spark、流计算、图计算、数据可视化, 以及大数据在互联网、生物医学和物流等各个领域的应用。在 Hadoop、HDFS、HBase、MapReduce、Spark 等重要章节安排了入门级的实践操作, 让读者更好地学习和掌握大数据关键技术。

2. 大数据软件安装和编程实践指南

详细学习如何安装、运行各种大数据软件, 以及如何进行初级编程实践, 包括 Hadoop、HDFS、HBase、MapReduce、Spark、MongoDB 等安装、操作、编程指南。

3. 备课指南

详细说明了教师如何备课, 包括教学大纲、讲义 PPT、授课视频、课后习题、上机题目等。

4. 授课视频

笔者主讲的全套大数据课程视频, 供老师上课参考。课程内容涵盖大数据、云计算和物联网概念及其相互关系、大数据处理架构 Hadoop、分布式文件系统 HDFS、分布式数据库 HBase、NoSQL 数据库、云数据库、分布式并行编程模型 MapReduce、图计算、流计算、基于内存的大数据处理框架 Spark、

基于 Hadoop 的数据仓库 Hive、大数据在不同领域的应用等。课程视频自 2016 年 3 月 28 日在网易云课堂正式上线以来，一直稳居热门课程榜单前列，深受广大网友的欢迎，国内多家知名大数据企业、慕课网站和培训机构采用本课程视频。截至 2016 年 12 月 12 日，网易云课堂学习人数突破 18000 人，收获评价中 99% 为五星级最高评价，被众多网友称为“国内难得的经典课程”。

5. 实验指南

用于机房统一上机，包含题目和答案。

6. Spark 入门教程

Spark 是当前最热门的大数据处理框架，笔者编著的《Spark 入门教程》，让初学者零基础零障碍学习 Spark。教程采用 Scala 语言编写 Spark 应用程序，因此，教程包括 Scala 入门和 Spark 入门两个部分的内容。

7. 大数据课程实验案例《网站用户购物行为分析》

采用 2000 万条用户购物数据集，案例涉及数据预处理、存储、查询和可视化分析等数据处理全流程所涉及的各种典型操作，涵盖 Linux、MySQL、Hadoop、HBase、Hive、Sqoop、R、Eclipse 等系统和软件的安装和使用方法。案例适合高校（高职）大数据教学，可以作为学生学习大数据课程后的综合实践案例。

四、结束语

随着大数据的全面普及，高校大数据专业建设也会加快推进，优秀大数据人才的培养和优质教学资源建设，离不开全国高校教育界同人，以及社会上的教育服务机构的共同努力。最后，祝愿我国高校大数据教学事业不断迈上新的台阶！

作者简介

林子雨：博士，现为厦门大学计算机科学系助理教授（始于 2009 年 7 月），海峡云计算与大数据应用研究中心副主任，曾任厦门大学信息科学与技术学

院院长助理、晋江市发展和改革委员会副局长。现为中国计算机学会数据库专委会委员，中国计算机学会信息系统专委会委员，厦门市计算机学会理事，海西创客学院首批 30 名创业导师之一，泉州市物流信息化产业技术创新战略联盟专家委员会副主任，厦门市物流协会信息化专业委员会委员。中国高校首个“数字教师”提出者和建设者，O2O 大数据教学理念提出者和践行者，厦门大学数据库实验室主要负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，厦门大学计算机系官方网站和厦门大学数据库实验室官方网站建设者和维护者。

运用大数据技术深化教育信息资源应用

成都信息工程大学软件工程学院院长 舒红平

2016年起,经研究,大数据应用是一个起点高、投入大、普惠难的场景。教育信息化从建设数字化校园进入信息技术与教育融合创新的阶段,为下一步建设智能化教育提供了支撑。

一、教育信息资源的定义

教育信息资源是指经过数字化处理,可以在计算机或网络环境下运行的多媒体信息材料。广义的教育信息资源是指在教学过程中学生和老师所接触、获得的经过加工处理的一切教育信息来源;狭义的教育信息资源是指以电子化、数字化、网络化为技术特征的教育信息资源。

二、数字化校园对教育信息资源应用存在“六大不足”

(一) 总体规划有度,实施路径不明

总体规划是面向未来的、方向性的。如何利用现有的教育信息数据、教学运行和管理数据、教育行业数据,通过实施路径分阶段、分批次实现规划目标,其路径是不够明确的。究其原因,是由于不具备对教育信息资源建设和应用的过程路径,效益评估缺乏。

对策:深化大数据应用,第一步是整合现有的多源数据,采用自上而下的数据思路,规划教学资源发展关键管控场景及指标,通过前置机、物联网设备等,建立指标对应观测点数据、常规和实时采集渠道。交给存储、加工、计算形成发展指标和汇总计算及整体态势感知,从而周期性或实时对教学质量进行监控。

通过建立教育信息资源大数据平台，以成都信息工程大学卓越人才、能力达成度和监控项目为例，集成整合教务、学生、教师、考试及就业等方面的数据，以学生成长成才目标进行各年级各阶段目标分解，将专业认证和审核性评估中的数据对应知识、能力及素质要求，以指标的方式分解到教学计划、教学大纲、课程及授课环节进行管理，学生学习效果对应知识、能力、素质达成等级进行动态评估和监管，从而有效提高教学质量。

图 1 所示为教育信息资源大数据平台的总体架构。

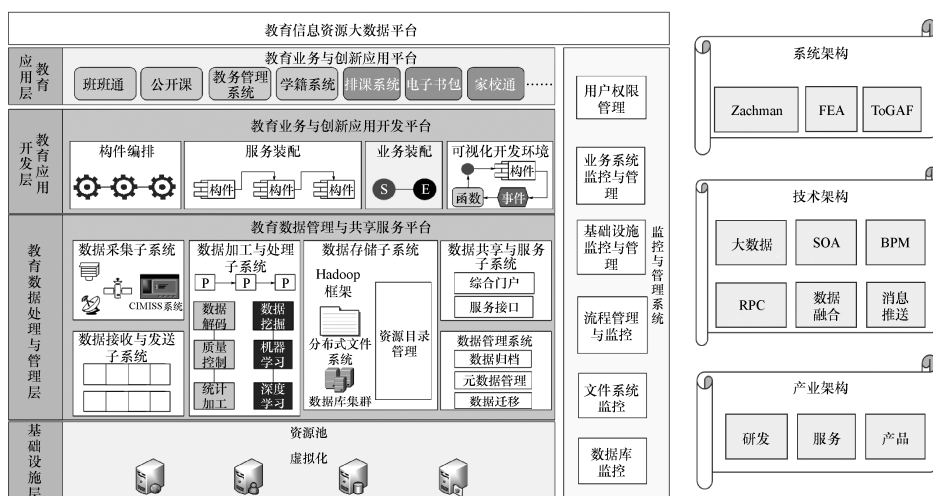


图 1 教育信息资源大数据平台的总体架构

（二）交汇循规蹈矩，独占滋生权杖

具体是指缺乏数据交换的平台和工具，使各个教学部门的业务数据难以及时分割、统一调度、有效共享，长期聚集的数据为滋生数据官僚提供了温床。

对策：建立统一的数据资源目录和交换平台，根据成熟的数据交换标准建立跨部门的数据交换共享服务，采用分布式数据存储实现对面向服务的部门数据资源共享和交换。以教育信息资源大数据平台交换的经验和策略性介绍，需要建立目录接口规范、教育信息资源分类规范、元数据标准、教育信息资源标识符、编目规范、目录体系技术管理规范、异构数据库接口规范等。资源目录需要建立信息资源主题目录、教育信息资源部门目录、教育信息资源

源交换目录，交换制度需要建立信息资源接入制度、资源共享奖励制度、信息资源查询制度、信息资源安全保密制度等相应的制度。

图 2 所示为基于可配的多种类数据源信息接入。

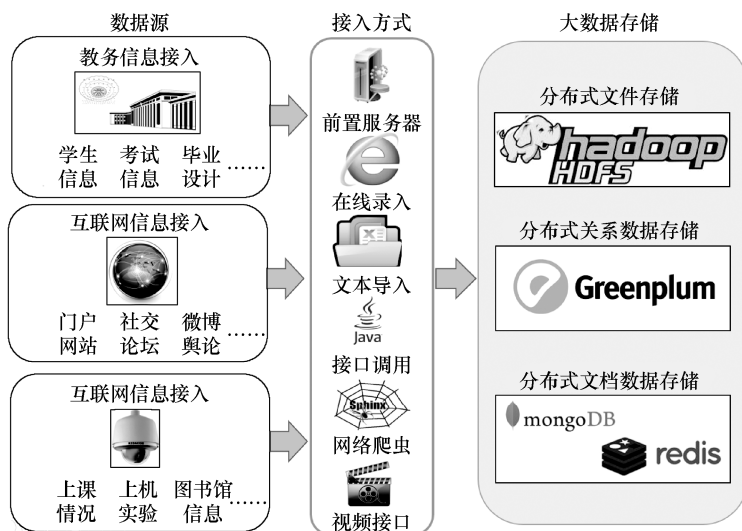


图 2 基于可配的多种类数据源信息接入

图 3 所示为建立标准的数据资源目录和交换平台。

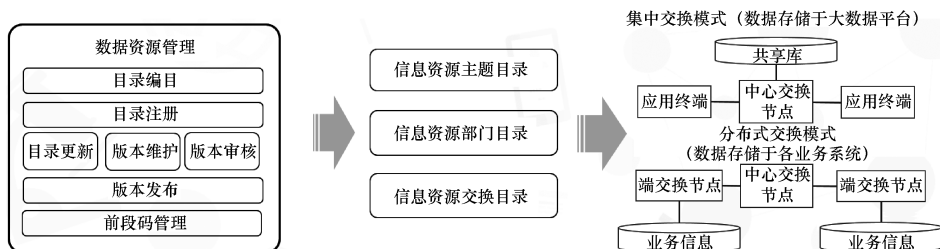


图 3 建立标准的数据资源目录和交换平台

（三）及时存储易行，整合计算难得

数字化校园面向职能的电子化和数字化，很好地解决了业务数据及时存储和信息资源灵活拓展问题，但在进行专题的全局数据业务分析时，因缺乏灵活适配的计算模型，对多元数据的重组能力、分布式计算量大、多个业务

计算模型组合难，成为应用起点高的突出表现。

对策：通过大数据开源技术可实现分布式计算，但关键是应用计算模型是一个持续改进和优化的过程，可以通过试探性地构建分析模型、分析模型如何快速适配、分析模型如何叠加和优化等问题，提供具有共性特征的领域模型来降低整合难度。例如，课程群的设置合理性和教学质量影响着因果关系分析：首先，判断课程群设置的合理性是否与教学质量存在因果关系；其次，当判断出教学质量变化是由课程群设置的合理性引起后，再进一步分析教学计划设计是否影响教学质量，主要包括课程群的科目设置是否合理、课程群的先后关系设置是否合理。通过对“Linux 体系和编程”这门课的教学质量分析，发现“操作系统原理”这门课学习比较好的学生其“Linux 体系和编程”的成绩也比较好。利用这个模型在“某课程的上课对象是不同学生群体时教学效果的变化”“某课程由不同教师群体担任造成的教学质量的变化”两个应用中进行快速适配、优化。

（四）专注部门精准，全局关联迷失

从学校全局角度讲，各部门和单元都有自己的一套数据分析工具或手段，单从某一视角，如从某个维度、线索对数据进行重构或再检索，甚至引入一系列数据的观测和采集，这必然引发追加 IT 投入，造成大的投入困境。

对策：全局专题分析既是大数据明显的优势，又是大数据前期应用的短板，维持有必要以专题应用为导向，向全局大数据架构分析平台发展，以支持众多以专题分析、规模化应用的角度提升 IT 应用的产出比。以学生跨画像专题为例，需强调专题应用开发的特点、专题应用如何引导专题数据架构、专题应用如何验证 IT 投入，学生画像专题应用通过数据整合、分析挖掘出每个学生的学习及生活状态，其中学习状态涵盖专业课、选修课、体育课、自习、借书等，生活状态涵盖吃饭时间、锻炼情况、经济情况、消费情况、同学关系等。建立学生在排名、预测方面的内容。

（五）人机查询丰富，学习机制简单

数字化校园中的系统大多是事务处理系统和管理信息系统，决策支持和管理系统少，系统的学习机制侧重于个性化服务。随着大数据应用的规模化

推广，尤其是对某群体资源包、历史数据进行策略提取时，学习机制变得复杂化。

对策：机器学习和可视化技术成为场景知识表达的利器，关注问题的表达和结构的可视化，将以教育资源的深度利用和体验为重点。以评教模型的优化为例，通过建立学生课程评教、问卷调查及同行评教的多模式评教方法，评教模型面临的决策冲突、个案冲突反映的模型缺陷，通过机器学习机制，规避这种缺陷引起对评价模型的不信任。

（六）数据挖掘有余，场景开发不足

以业务为中心的数字化校园建设很重视对历史数据的挖掘分析，但这些分析多侧重于决策参考与知识发现，数据多来自数据局势、多维数据分析角度。随着教育科学化、规模化管理的日益增强，教育管理的决策点也在不断增加和更替，动态提供面向不同决策场景的支持工具。

对策：提供众创微服务开发工具和平台，使开发成为解释和洞察数据的工作。以互联开放大数据众创平台经验为例，需做以下几方面的工作：通过微服务提供微决策的支持；微服务开发中的构件化、可视化、服务化开发模式，代码自动生成、实施加载动态运行；微开发中场景的应用、用户的参与都是需要关注的，如图4所示。

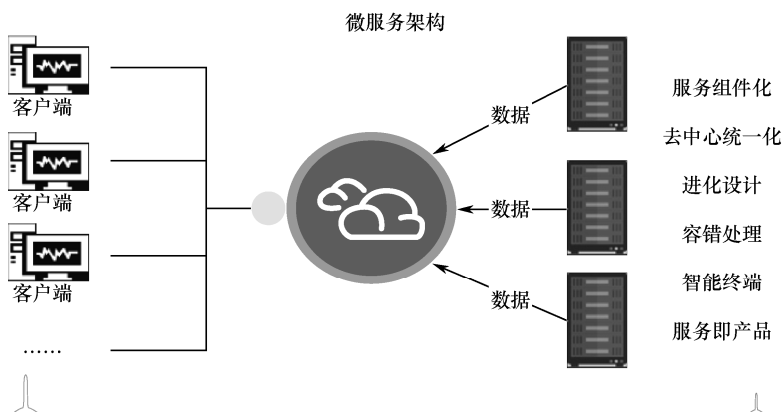


图4 微服务平台架构

三、总结

深化教育信息资源的应用需要在深化管控、深化交汇、深化模型、深化专题、深化交互、深化规模上进行教育资源的应用。其中在深化管控方面进行统一规划、分步实施，标准化执行；在深化交汇方面建立统一的数据资源目录和交换平台；在深化模型方面重视共性特征的领域模型；在深化专题方面以专题应用方式引导大数据平台总体架构；在深化交互方面关注于问题表达和结果可视化；在深化规模方面以微服务开发方式提供微决策支持。

作者简介

舒红平：CIO 时代学院第 22 届 CIO 班学员、成都信息工程大学软件工程学院院长，四川省学术与技术带头人后备人选，国家气象通信与信息技术委员会副主任委员，气象信息共享与数据挖掘四川省高校重点实验室主任，智能信息处理与制造业信息化研究所所长，四川省科技青年联合会委员，四川省、成都市科技计划评审专家，四川省创新企业评审专家。成都信息工程大学软件工程学院学术科研骨干及计算机应用技术省级重点学科数据库及知识工程学术带头人，努力学习科学文化知识，治学严谨，学科方向明确，一直致力于数据库技术、计算机在制造业中的应用、数据融合及数据挖掘方向的研究。

大数据在高校智慧校园中的应用

上海科技大学图书信息中心总工程师 孙名松

一、小数据时代与大数据时代

“数据(Data)”在拉丁文里的意思是“已知”，也可以理解为“存在”。所以，“数据”就是“存在”，“大数据”就是“大存在”。研究大数据，就是研究大存在，也即研究一切物质、一切行为、一切思想，以及人类自身^[1]。

数据充斥并改造着人们的生活、工作。数据化是指把现象转变为可指标分析的量化形式的过程，其中包含对世界的梳理、理解，并形成可保存的经验。计算和记录共同促成了数据的产生，是数据化的根基。而数字化是把模拟数据转换成0、1表示的二进制码，方便人类使用现代技术对数据进行更好的处理。数据化是一种思想，数字化是一种手段；数据化古而有之，数字化方兴未艾。

小数据时代依靠随机采样，其原则是以最少的数据获得最多的信息。但如此，则无法了解一些微观细节，不利于对某些特定子类进行分析。而“参差不齐是世界的本质”，细节缺失将会影响对整个自然活动、人类活动的探索与研究。此外，随机采样以研究者的理论前提为设计基础，只能对已遴选的问题进行解答，而难以虑及其他问题。也就是说，小数据时代是以极其有限的信息面对有“偏见”的问题。

大数据时代，意味着将世界数据化，意味着世界的本质就是信息。世界不仅被看成一串事件的组合，更被看作信息的集合、数据的集合。这是世界观的深刻变革：人类具备以往认识并处理事件的经验而不盲从于经验，人类采集“数据”，但更明确“所见、所思、所得”皆为“数据”，我们生活在数据的海洋之中。

以上，从小数据时代到大数据时代，伴随或产生了以下几种转变与认识：

- (1) 意识到“样本”等于总体。用更大、更全、更综合的态度来观察、理解、关照世界。
- (2) 大数据对于精确性的要求降低。在小数据时代，因为数据少，所以，对数据的精确度要求非常高，而当大量数据出现时或者要求数据量大时，必然需要接受数据的纷繁复杂。
- (3) 要意识到数据错误并不是大数据的固有特性，而是需要处理的实际问题，该问题可能长期存在。
- (4) 混杂绝不等于错误。混杂是大数据的常态，且应该是一种基本态和标准态。
- (5) 大数据揭示了传统样本无法揭示的细节信息，大数据是通往“精准”处理的基本途径。
- (6) 大数据时代，不再热衷于追求因果关系，而是试图探寻不同事物之间的关系，在此基础上找到可供观察的关联物，以进行预测。而预测，是大数据应用的核心所在。
- (7) 相关关系被阐释之后，可进行因果关系的分析。但是必须注意到，因果关系只是相关关系的特殊形式，因果关系在大数据时代已经不是解释世界的基础；相关关系是一种较为普通的存在，在大数据时代更容易被发掘，可以更高效地指导实践，随着大数据的发展，以往的因果关系可能会被证伪，或被视为相关关系。

其中第(1)点是大数据对于认识论的改造；第(2)～(5)点体现了大数据时代与传统时代对数据要求的迥然不同；第(6)和(7)点则是数据间逻辑关系的优先性的颠覆。从实践的角度而言，第(1)点可以作为前提，第(2)～(5)点可以作为数据搜集与处理的准则，第(6)和(7)点或可作为数据解释的指导方向。

二、大数据在高校智慧校园中的应用

关于大数据如何定义，研究机构 Gartner 的定义如下：大数据是指需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。麦肯锡的定义如下：大数据是指无法在一定时间

内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合。舍恩伯格·维克托的《大数据时代》中的定义如下：大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法。

北京理工大学张华平副教授给出的定义如下：大数据是指从客观存在的全量超大规模、多源异构、实时变化的微观数据中，利用自然语言处理、信息检索、机器学习等技术抽取知识，转化为智慧的方法学^[2]。

无论哪种定义，我们可以看出，大数据并不是一种新的产品，也不是一种新的技术，就如同 21 世纪初提出的“海量数据”的概念一样，大数据只是数字化时代出现的一种现象。那么，海量数据与大数据的差别何在？从翻译的角度来看，“大数据”和“海量数据”均来自英文，“Big Data”翻译为“大数据”，“Large-scale Data”翻译为“大规模数据”，“Very Large Data”翻译为“超大规模数据”，“Massive Data”则翻译为“海量数据”。从组成的角度来看，海量数据包括结构化和半结构化的交易数据，而大数据除此以外还包括非结构化数据和交互数据。Informatica 中国区首席产品顾问但斌进一步指出，大数据意味着包括交易和交互数据集在内的所有数据集，其规模或复杂程度超出了常用技术，按照合理的成本和时限捕捉、管理及处理这些数据集的能力。可见，大数据由海量交易数据、海量交互数据和海量数据处理三大主要技术趋势汇聚而成。

（一）大数据的特征

大数据的特征包含四个层面。第一，数据体量巨大。从 TB 级别跃升到 PB 级别。第二，数据类型繁多。例如，网络日志、视频、图片、地理位置信息等。第三，价值密度低。以视频为例，在连续不间断的监控过程中，可能有用的数据仅仅有一两秒。第四，处理速度快。最后这一点和传统的数据挖掘技术有着本质的不同。业界将大数据的特征归纳为 4 个“V”，即 Volume、Variety、Value、Velocity。

1. 数据体量巨大（Volume）

大数据通常指 10TB（1TB=1024GB）规模以上的数据量。之所以产生如此巨大的数据量，一是由于各种仪器的使用，使我们能够感知到更多的事物，这些事物的部分甚至全部数据就可以被存储；二是由于通信工具的使用，使

人们能够全时段联系，机器—机器（M2M）方式的出现，使得交流的数据量成倍增长；三是由于集成电路价格降低，很多东西都有了智能的成分。

2. 数据种类繁多（Variety）

随着传感器种类的增多及智能设备、社交网络等的流行，数据类型也变得更加复杂，不仅包括传统的关系数据类型，也包括以网页、视频、音频、E-mail、文档等形式存在的未加工的、半结构化的和非结构化的数据。

3. 价值密度低（Value）

数据量呈指数增长的同时，隐藏在海量数据中的有用信息却没有以相应比例增长，反而使我们获取有用信息的难度加大。

4. 流动速度快（Velocity）

我们通常理解的数据流动速度是指数据获取、存储及挖掘有效信息的速度，由于现在处理的数据是 PB 级代替了 TB 级，“超大规模数据”和“海量数据”也有规模大的特点，数据是快速动态变化的，因此，形成流式数据是大数据的重要特征，数据流动的速度快到难以用传统的系统去处理。

大数据的“4V”特征表明其不仅仅是数据海量，对于大数据的分析将更加复杂、更追求速度、更注重实效。

大数据独立的发展形成特有的市场化与规模化，也充分带动了其他行业与大数据的广泛、充分融合，从而推进了大数据的全面落地。大数据从产业到行业的成熟将推动更多传统企业向科技智能化转型，将推进政府政务大数据发展，也将鞭策大数据行业在中国平稳落地，与其他各行业共襄中华民族伟大复兴之盛举。大数据之大，还在于数据结构的有容乃大，它不再需要传统的数据库表格来整齐排列，几乎可以无所不包地记录、存储和计算各种规则的结构化数据和不规则的非结构化数据，于是便有了逐步演变为一个数字化世界的可能。

2015 年国家提出并制定了“互联网+”行动计划，将“互联网+”上升到了国家战略。“互联网+”的提出必将给高校智慧校园建设增加新的内涵、注入新的动力。借助“互联网+”推动数字校园加速向智慧校园升级，充分利用云计算、物联网、移动互联、大数据等一系列新技术、新理念、新模式，打造全新的大学智慧校园，有力支撑大学未来发展战略，带动人才培养及评价

方式的创新、提升校务治理水平，提供多层次的个性化服务和智能化管理决策，大学智慧校园建设的核心内涵可以概括为“全面的环境感知、无缝的网络互通、弹性的云生态圈、海量的数据支撑、开放的学习环境、个性化师生服务、智能化管理决策、高效的校务治理”。

2015年8月，国务院印发了《促进大数据发展行动纲要》。纲要明确指出，要将大数据技术应用于我国的文化教育领域，建立教育文化大数据。从目前全国各高校的信息化规划来看，都将大数据列为重点工作内容。

高校在信息化进程中，产生了各类结构化、半结构化和非结构化的数据，包括教学管理数据、教学资源数据、学生信息数据等，大到高校的治校方针政策，小到学生的日常消费，数据繁多，类型复杂。然而，高校要想发挥这些数据的潜在价值，探索教育大数据的应用，面临以下挑战：数据接口不完善，有效数据质量不高；对信息化软硬件和运维服务要求高；建设需要关联分析、统一规划；高校的用户类型多，需求各异。

利用大数据技术对这些数据进行收集、分析，转化为高校管理与服务可利用的资源，将对智慧校园建设起到非常重要的作用。

（二）举例说明大数据技术在智慧校园中的应用

1. 综合校情展示

对学校管理者而言，通过综合校情分析展示，可以对学校的在校生情况（本科生、研究生）、课程情况、科研成果情况、奖助情况、就业情况、教工情况、教师分布、干部情况、家具情况、资产情况、房屋情况、排名情况、消费情况等方面进行直观的了解和横向及纵向的对比。结合历年数据变化规律可以为辅助决策提供依据。不同系统之间数据的关联性或许能够给管理者决策提供新的思路。

综合校情展示主要包括基础数据分析展示和行为数据分析展示。

基本数据分析：如招生数据分析、学生数据分析、毕业数据分析、教师数据分析、课程数据分析、成绩数据分析、就业数据分析、高校资产数据分析等。

行为数据分析：学校食堂就餐情况分析、一卡通消费行为分析、上网行为分析、图书借阅行为分析、图书馆使用时长、上网时长/流量和成绩之间的

相关性分析、重点人群群体的特征刻画分析和预警等。

举例说明：

(1) 高校就业信息统计。从高校学生的毕业去向、就业单位、就业地区、就业行业、就业薪资等多维度进行统计分析，全面呈现高校就业情况，为高校就业办发现学生就业规律、有针对性地进行学生就业指导提供支撑。

(2) 教学信息统计分析。为校领导呈现了高校热门课程排行、各院系开设课程统计和学生成绩统计分析、挂科率分析，全面呈现学生在校期间的学习与成绩分布，为指导高校课程开设、提高学生成绩提供支撑。

(3) 一卡通统计分析。展现了高校学生整体消费能力、消费偏好，为后勤部门了解学生餐饮、购物偏好，有针对性地进行提升服务水平提供支撑。

(4) 各生源地消费能力。按照生源地统计该地区学生的消费能力，详细查看在某一段时间学生消费额和消费次数的统计。

(5) 学校网络使用状况分析和学生上网行为统计。通过对学生上网的地址进行统计、分析，结合其基础的个人信息数据，可按不同的维度，如性别、籍贯、院系等来统计出不同类别的人群。对于某类网站的使用频率，如果记录的日志足够详细，甚至可以统计出学生在网上消费的喜好或偏向，对于后勤或学工等部门也是一个比较重要的参考。

应用到的相关技术有数据关联分析、多源数据整合、海量日志数据处理、benchmark、指标体系建立、Agile BI、全文检索引擎。

2. 公共资源使用情况分析

对于高校而言，食堂、体育场馆、教室、图书馆、校医院等各类公共资源有限，师生没有很好的途径获知这些资源的服务能力情况，导致经常发生排队、拥挤的情况，给师生学习、生活带来了不好的体验。随着学校信息化的推进，各部门管理信息系统逐步建设并投入使用；随着技术的发展，特别是物联网和智能感知设备的出现，使数字校园智能服务成为可能。

数据来源于一卡通消费、一卡通门禁、无线网、校园安全视频监控等。

(1) 食堂、澡堂人员密度状况及建议各食堂、公共澡堂各时段就餐人员密度情况，各类人员（年级、籍贯、职称等）就餐爱好、习惯等。

(2) 教室使用状况、人员密度、各时间段教室使用情况、教室人数等；

基于无线网络进行考勤。

(3) 会议场馆、体育场馆使用状况及人员密度。为师生提供会议场馆的可用性查询, 体育场馆的使用情况(有课、无课等), 以及人员密度发布。

(4) 图书馆座位使用状况及人员密度发布, 提供图书馆座位空闲情况及图书馆内人数等。

(5) 校内人员密度分布。根据学校无线网数据、安全视频监控信息, 识别学校人员热力分布图。

应用到的相关技术有数据关联分析、数据挖掘(聚类分析)、海量日志数据处理、多源数据整合(日志数据与结构化数据整合)、高速内存数据库、分布式全文搜索引擎。

3. 个人数据报告

面向校园师生用户提供个性化数据服务, 展现师生在校园内学习、消费、生活、健康等方面的个人行为习惯, 通过严谨的数据分析帮助学生更加了解自己, 以及与他人的差异, 帮助校园师生感受信息化带来的人文关怀与改变。

数据源自一卡通消费、图书馆门禁、图书借阅系统、校园网络系统、体育场馆门禁等, 形成的报告如:

- (1) 校园卡账单及消费习惯分析报告。
- (2) 图书馆进出频次、时长及借阅习惯分析报告。
- (3) 网络账单及上网习惯分析报告。
- (4) 体育健身锻炼学期报告。

这些报告可通过高校官方微信号、APP 进行手机推送, 移动互联网时代方便用户及时阅读、分享、传播。

应用到的相关技术有数据关联分析、数据挖掘(用户画像)、海量日志数据处理、多源数据整合。

4. 图书馆电子期刊资源使用效率分析

高校每年花费资金购买著名期刊论文集, 为师生用户提供便捷的文献检索和下载服务。图书馆电子期刊资源的使用情况、不同学科对于不同电子期刊资源使用偏好的差异, 是图书馆亟须了解的内容。通过对高校用户期刊文献检索记录的大数据分析, 优化论文期刊购买方案, 使图书馆可以采购到师

生更加需要的资源（传统纸质+电子资源），提高现有采购效率。

学校通常的做法是向数据商（如万方、CNKI）购买电子期刊资源访问统计数据，而这种方式基于学校整体访问数据做统计分析，无法基于用户做访问详情的分析统计，从而无法获取到基于不同学科门类、不同学院和专业特点、不同教师等级的不同人群期刊访问情况分析，也无法了解到不同资源库的使用情况横向对比分析。对师生的检索关键词进行挖掘也是非常重要的方向，而传统的做法无法了解学校师生用户检索电子期刊资源的检索偏好、检索热门等具体信息。

出口网络日志数据记录了师生访问电子期刊资源库的行为，通过大数据技术对出口 URL 日志等数据进行处理及关键信息提取，关联学校内部用户信息数据，将实现图书馆电子资源使用的全面分析及人群分析，为图书馆采购决策提供辅助。

数据来源于图书馆采购电子期刊资源列表、师生上网 URL 日志、师生上网身份认证等。

应用到的相关技术有数据关联分析、海量日志数据处理、多源数据整合（日志数据与结构化数据整合）、分布式全文检索引擎。

5. 校园舆情监测

在移动互联网大潮之下，无论是正面信息还是负面信息，都会以更快的速度传播。学校声誉对学校招生、就业、评优评先等方面有很大的影响，随着移动互联网和社交媒体的普及，高校越来越重视学校的社会评价。目前部分高校会利用互联网数据监测学校声誉，通过大数据的手段实时监测互联网新媒体上与学校相关的新闻、传播话题和用户反馈，了解学校舆情、声誉及影响力。

应用到的相关技术有文本挖掘、语义分析（正、负面判断）、语义相似度计算、弹性爬虫引擎、分布式全文检索引擎。

大数据在智慧校园中的应用还包括教学信息统计分析，通过对课程知识结构进行样本分析，结合教育过程，综合学生学习成绩分布来验证课程讲授过程的合理性和工程教育认证中的达成度，来综合分析课程开设的合理性。

又如，学校资产管理信息分析，借助于资产管理信息平台实现对校园基

基础设施、教学实验设备、校园通信网络设备等数据的采集分析，为学校基础设施建设方向、教学实验设备的维护、校园网通信设备的升级改造提供数据支持。

参考文献

- [1] 孙九林, 任博. 数据与科学大数据, 中国科研信息化蓝皮书 2015[M]. 北京: 科学出版社, 2016.
- [2] 张华平, 商建云, 段永朝, 等. 大数据大家谈[M]. 北京: 中国工信出版社, 2017.

作者简介

孙名松: 现任上海科技大学图书信息中心总工程师。曾任哈尔滨理工大学网络信息中心主任, 哈尔滨理工大学软件学院院长、教授、硕士生导师。
研究方向: 高校信息化建设、网络信息应用、网络信息安全。

区块链+教育的需求分析与技术框架

岭南师范学院网络与信息技术中心主任 金义富

一、区块链的概念理解

我们每天的工作生活会产生大量的数据，而我们却无法掌控这些数据。这些数据被第三方（如银行、通信运营商等）所拥有，我们的一举一动都可能被第三方平台记录，这些数据可能侵犯我们的隐私，同时也会为数据持有方带来一些利益。在教育领域也是如此，教育云平台、资源中心等也记录着我们的学习状态，我们也未得到很多的反馈。区块链技术的出现将改变这种状况。

区块链的组织结构，其实是由区块按照时间顺序组成的一条链，每一个数据区块包括区块头和区块体两部分，接着由一些算法进行封装。区块链一个很重要的特点是去中心化和自信任，它的数据采用分布式存储，每个节点都能复制一份完整的数据库副本，按照过半数同意才有效的原则，如个别的或是一般少于 50% 的数据修改对区块链整个的数据均无影响。因此，数据区块不太可能被伪造。同时，数据区块还带有时间戳功能，可以具备数据溯源的功能。区块链的这种结构最大的特点是去中心化，比如我们通过淘宝向商家下订单，钱打到第三方，确认之后由第三方向商家支付，这是一种间接的形式，此时的第三方则显得尤为重要，如果第三方出了问题，这个问题是很大的。区块链技术其实可以去掉第三方，实现买卖双方的直接交易。

关于智能合约，是区块链系统第二代的一个重要功能。例如，张三和李四打赌明天甲乙两队比赛谁赢的问题，如果甲队赢，张三给李四 50 元；如果乙队赢，李四给张三 50 元，为保证其顺利执行，可能有如下 3 种方式供选择：第一种是双方定好协议，在结果出来后执行，如果不执行就起诉到某个机构；第二种方式是每人交 50 元给第三方王五，由王五来执行合约，如果王五跑了，

也是有风险的；第三种方式便是智能合约，利用电子货币的形式通过某种程序来实现控制，在比赛结束后，这个程序会自动进行判断、扣款。建立物联网上的智能合约可以实现物体与物体之间的对话，比如家里的电冰箱盛放了冰激凌，假设电冰箱本身发现冰激凌数量不足，电冰箱自己可以向商家下订单进行购买，实现冰箱自动的智能购物。

总之，区块链可以重构全新的信用体系，降低欺诈风险，优化组织结构和业务流程，提升资源品质和共享效率。随着物联网、人工智能和大数据等技术发展和区块链应用领域的不断增加，有望形成类似于今天互联网一样覆盖全领域的价值互联网，所有人和机器都可以连接到一个统一的全球互联网中。

二、区块链的教育需求

（一）教学资源

区块链的去中心化和共维护等特点，可以实现教学资源的定制共享，教师和学生都可参与这种资源的更新，让学生自己生成资源，通过集体的维护、集体的智慧使这种资源得到最大程度的优化。同时，区块链可以保证资源的信任度和知识产权。可以利用区块链技术支撑分享式自主学习，每个学习者都可以成为一间微学校，这种分享式学习是在奉献中的学习，在教授别人过程中的学习，而教师也在此过程中实现了自身水平的提升，因为整个区块链的资源是由师生共同维护的。

（二）教学评价

在教育中第二种形式的区块链应用，从教学过程和评价来看，也是区块链在学习教育领域中目前已有实现的案例。如建立学习者学习记录的区块链，我们设想，可以将一节课作为数据的区块，信息来源于参与课堂讨论与测试的大数据系统，另一种形式来源于自学或非正式学习的记录，以一个项目或者一个任务组件数据区块。其实在任意的时间，不论学习者在哪里，包括机器人的学习，如果这种区块链的评价体系、记录形式建立后，过程中的数据可靠性与区块链本身固有的特点是对应的，那么可靠性极高。目前国外已有大学将这种形式的学习评价直接作为学生晋级、招生，甚至毕业之后学历认

证的个人评价依据。

（三）教育培训与支持系统

教育培训与支持系统也比较明显，如建立全国的区块链系统，实现在职教育、培训、教育扶贫，特别是教育扶贫，因为与经济直接相关，对投入产出进行非常客观的、合理的、区块链式的记录。

（四）学校内部管理

通过整合学校的 OA、人事、教务、科研、财务等系统，实现学校的数据管理，如建立学校自己的财务收支区块链，由项目经费来源单位与经费负责人直接相互信任，不再依托于财务，让来往的账目计入交易的区块，由区块链固有的共识制确认数据的安全，如科研经费的报销可能会发生一些新的变化。可以预测，在未来，学校管理这部分应该可以形成依托于区块链的智慧校园，实现教学信息互用和师生的价值互联，这一块前景非常广阔。

三、区块链+教育的技术框架

区块链+教育的技术框架的主要思路是将整个教育系统抽象为人与教育资源两个最重要的因素，通过人和教育资源发生的一系列关系来建立数据区块，形成一个链，人包括教师、学生教育管理者，以及教育活动中的所有人。教育资源是所有的教学活动汇总存在的有形和无形的各种资源的统称。总体上，将整个教学教育活动抽象成两大最重要的因素：人与教育资源，通过它们之间的相互作用来形成一个数据结构。

区块链+教育的总体结构并未抛弃目前流行的教育云中心结构，在优质资源共享平台的教育云中心还是会持续存在的。总体结构设计为一种部分去中心化的混合部署的模式，云平台是中心化的架构，交易区块链是对等网络的架构，二者互为依托，将优质资源放到云中心，分享资源通过区块链进行组织，其中还可进行集中维护与集体维护互为补充。有关教育区块链的对等网络、开源编程可充分体现以学生为中心的教育理念，可以实现本地化的数据管理，可以实现数据隐私保护，可以做到我的数据我做主，从技术本身激发

学生参与的积极性。按照上述结构，我们团队已进行了初步的研发。

目前区块链技术的发展刚刚起步，可以说在教育领域中的应用还未起步，但教育改革对区块链的需求非常迫切。

作者简介

金义富：现任岭南师范学院网络与信息技术中心主任，广东高校数字化学习工程技术开发中心主任、博士、教授，曾任岭南师范学院信息学院院长、科研处处长。他是中国人工智能学会计算机辅助教育专委会常务理事，广东省计算机学会软件工程专委会副主任委员，湛江市教育信息技术协会理事长。主要研究领域为数据挖掘与教育信息化，“网真课堂”成果获得湛江市科技进步二等奖。主持完成省部级以上课题 5 项，发表论文 70 余篇，获得专利 10 余项。

本科大数据实验平台及资源建设等 的思考与探索

中国农业大学信息与电气工程学院博士 李 辉

对农业而言，大数据既是机遇，又是挑战，只有挑战大数据，使信息技术处于农业领域的制高点，才能充分发挥大数据的优势，从而为农业助力。

一、关于农业大数据的认识

农业大数据是指以大数据分析为基础，运用大数据的理念、技术与方法处理农业生产、销售整个链条中所产生的大量数据，从中得到有用的信息以指导农业生产、经营、农业流通和消费的过程。农业数据应用作为农业大数据产业的落地点，分析挖掘数据的价值，还原大数据结论，反映行业问题。换言之，将农业大数据应用于粮食安全、土地经营、病患防治、动植物育种、农业结构调整、农产品价格、农副产品消费等领域，解决农业生产过程中遇到的问题。但农业数据是很复杂的，具体表现为数据源分布广、可控制度低、作物干扰大、类型多样、结构复杂和获取困难等，这些因素导致我国农业大数据面临着诸多挑战和问题：首先，大数据研究普遍存在只有数据、没有充分应用价值的问题，导致收集数据、存储数据的收益成本比很低；其次，数据类型单一，只有结构化数据，半结构化、非结构化数据的缺失导致数据的不完整。同时，也缺乏农业现代化与信息化的深度融合，区域视角缺乏全国视角；最后，基础数据采用业界的 Hadoop 开源技术简单堆砌，很难保证未来的实用性。

目前，专门从事数据科学与应用研究的人才比较紧缺，大数据人才的招募、培养、使用是农业大数据研究面临的最大挑战。因此，大数据产业的发展对大数据人才提出了新的需求，国内各高校陆续进行大数据学术研究的同

时，也在考虑将大数据相关课程纳入培养体系，以满足社会对大数据人才的需要。在我国，除了以山东农业大学为首的农业大数据产业技术创业联盟之外，还有江苏、中科院大数据实验室等陆续成立。中国农业大学作为中国农业院校的领军者，在大数据领域集中圈地建立农业大数据实验室。同时，在全国农业领域的积累及联合全国优秀的企业共同建立农业大数据实验室，并将农业大数据实验室教学尽快纳入培养体系之中，确保中国农业大学在农业大数据领域后来居上，达到国家级大数据重点实验室和农业大数据领域的领先地位。

在人才培养中，结合农业行业的相关应用特点的实验室教学是关键环节，满足农业行业的人才技能要求，需在本科的相关学科中强化基于农业行业相关数据的实验教学环节。首先要立足于信息与电气工程专业，面向全校本科生开展双学位大数据教学，从验证性、实际性和创新性 3 个层次设置实验，确保中国农业大学各个专业的学生可以通过此课程，了解大数据发展的新趋势和新动向，以及其对现代农业的影响和意义。其次，信息技术的发展为丰富教学手段提供了可能，通过开放共享大数据实验室资源，以联合大数据的科研院所开展农业大数据教学科研工作，全面提高整个农业院校的科技现代化教育水平。综上所述，为培养大数据教育的高新技术加农业相关分析技术结合的教育是一个大的尝试，从而为中国的农业现代化与信息化的快速发展培养合格的后备人才。

二、农业大数据本科实验室教育建设的目标

按照中国农业大学厚基础、宽口径、重实践、重交叉学科的要求，科学设立大数据人才培养方案，既要熟悉数据分析，又针对相关业务的不同要求，开始酝酿本科的农业大数据实验室教学设置，可以让学生了解农业大数据分析技术原理和实验方式，掌握大数据对农业相关专业所能带来的帮助及变革。为此，要达成以下 4 个建设目标。

目标一：建设业界领先的农业大数据实验室。结合中国农业大学在农业领域的丰厚积累和宝贵资源，以大数据技术与应用概论这门学科为公共课，使各专业本科均能受益，成为各大院校农业大数据实验室建设的样本，进而成为国家级农业大数据重点实验室。

目标二：建设融合农业行业经验、业界最新技术、科研教学实践与业界实际案例同时运行的新一代农业大数据实验室教育平台。此平台的建设不是一蹴而就的，是随着大数据技术的发展及农业实际案例的不断发展迭代更新，保证教学内容与时俱进，最大限度地避免传统教育知识陈旧，为农业现代化、信息化与先进技术的接轨和同步奠定基础。

目标三：运用先进的“互联网+”教育的线上线下相结合的教学模式，进一步扩大农业大数据实验室的覆盖范围。其中的重点是在上述农业大数据实验室平台上建立相应的大数据技术与应用概论课程，包括实际案例教学材料、教师教案实验用书、学生案例实验用书、答疑等相关教学工具与教学辅助材料。

目标四：面向学生就业和社会既定需求为前提的方向转变。针对农业经济、农业气象、生物信息、食品营养、食品安全、食品风险监测等专业的实际案例，实现大数据行业应用范例教学材料，可以考虑分期实施并根据需要进一步扩充和优化。

三、农业大数据本科实验室建设的可行性

为达成面向本科的大数据实验课程目标，中国农业大学采用了业界先进的平台和贴近农业实际的相关案例分析，充分考虑技术和专业的融合，从而保证课程的可行性和有效性。考虑到中国农业大学除信息与电气工程专业的学生之外，其他学院的学生并非为农业大数据相关专业，因此，课程的设计过程必须考虑广泛的实用性，于是将其细分为农业大数据技术应用课程与农业大数据创新与开发课程。

第一，学校师资资源的充裕保证。因为中国农业大学已建立了数据科学研究中心。同时，基于信息与电气工程学院的师资资源可充分保证农业大数据实验室的授课资料。

第二，本科生乐于拥抱大数据技术。以中国农业大学之前开设的大数据选修课基本情况反馈来看，本科生普遍热衷于学习新的技术，并运用新的技术解决新的问题。无论是校内大数据科研中心还是外部企业，对大数据的巨大人才缺口都是潜在的要求。

第三，校企联合可保证大数据实验室的先进性和实用性。实验室的搭建

与农业案例的开发可以联合业界优秀企业保证其先进性、可靠性、实用性，同时通过后期服务不断进行升级，保证技术不断更新与同步。

四、农业大数据本科实验室建设的方案

基于农业大数据实验室的建设目标与可行性分析，提出的建设方案主要包括农业大数据源数据、农业大数据实验室硬件平台、农业大数据实验室软件平台、农业大数据实验室资源的开发、大数据实验平台设计方案和大数据实验室案例教学的开发 6 个方面的内容。

第一，农业大数据源数据。数据是大数据分析的基础，主要包含农业经济、农业迹象、生物信息等。数据来源多种多样，数据类型除来自各个应用系统传统意义上的结构化数据、半结构化数据外，更多的是非结构化数据源，这些是大数据平台的原材料，我们将其称为“裸数据”。

第二，农业大数据实验室硬件平台。大数据的分析必须有硬件平台作支撑，农业大数据实验室硬件平台包括服务器、存储设备、网络投影仪和大屏幕等硬件，这些是大数据软件平台的定性基础。

第三，农业大数据实验室软件平台。农业大数据实验室平台主要包括数据准备、数据处理、数据建模和展现等软件平台。在数据准备方面，与商业智能类似，如果数据需要通过大数据平台进行处理，数据的前期准备工作显得尤为重要，如数据的抽取、清洗、转换和加载，相当于对于原材料进行粗加工，以便为大数据处理做好充分的前期准备。在数据处理方面，主要用 ETL 工具准备好数据，首先存储到分布式文件系统中，利用一系列商务智能分析对结构化数据进行分析 and 处理，进而达到数据挖掘和价值发现的目的，这是实现数据变为有效信息的第一步。数据的建模与展现方面，结果数据处理后，数据的价值可以通过进一步的建模工具、可视化工具从不同应用进行深度数据挖掘、决策支持等工作，让大数据针对某一行业或应用场景进行二次开发，这一步是信息变为知识的关键一步。

第四，农业大数据实验室资源的开发。数据实践运营中都存在开发实践教材资源，农业大数据实验室资源的开发主要是教学案例的开发，包括农业相关专业的数据模型开发、数据可视化等内容，农业大数据实验室资源的开

发主要包括大数据应用实验教程、大数据实际开发实验课程，农业大数据实验课程幕后的开发将知识供学生立即使用，农业大数据实验室教材的开发主要针对于教室和学生两个方面进行编写。

第五，大数据实验平台设计方案。大数据实验室平台的搭建以多节点、集群服务器平台作为农业大数据建设的硬件平台，提供大数据机能和分布式存储平台，以 Hadoop 作为应用的分布式存储平台，这个平台可以形成六大优势：第一，一个平台可以覆盖从数据到信息、从知识到支配全生命周期的流程；第二，数据模型可支持 99% 的数据接口，降低数据的接口难度，同时可支持结构化和非结构化的数据类型；第三，高度提升系统，无须面对多种系统混杂在一起带来的管理或复杂难题；第四，高性能，采用共享式文件系统，大大提高了数据处理和分析速度；第五，采用对等架构，降低故障风险；第六，基于 GUI 的管理模式，大大降低了管理难度。

第六，大数据实验室案例教学的开发。为让学生真正了解大数据在农业相关专业具体的应用场景和关键作用，有着感性和理性的认识，大数据作为当今的先进技术，对传统农业的升级有着巨大的推动作用，可以激发学生的学习热情，提高学生的创新思维能力，从而为学生将来的就业及推动农业现代化和信息化的发展奠定坚实的基础。为此，在大数据实验平台，针对农业专业方向提供相关的案例也是很重要的，如农业经济大数据案例、农业气象大数据案例、农业生物信息大数据案例等。对于案例的设计和研发，建议本着忠于实际、分步实施的原则应用于教学，在教学反思中快速迭代后续的案例开发。从而始终保证案例的真实性、可用性和有效性，更好地将大数据理论、实验平台与案例相结合，从而达到学以致用目的。

随着实验教学的展开，越来越多的学生会对新技术产生兴趣，从而成为我国农业教育的领航者，打开教育改革的新篇章，为大数据教学提供很好的思路，加快大数据在我国支柱产业农业方面的应用和发展。

作者简介

李辉：博士，中国农业大学信息与电气工程学院院长助理，农业大数据实验室主任。主要从事大数据技术在农业应用中的研究工作。

民族教育信息化建设探索与研究

云南师范大学民族教育信息化教育部重点实验室教授 甘健侯

一、民族教育信息化教育部重点实验室

（一）实验室概述

以云南师范大学为依托单位，民族教育信息化教育部重点实验室于 2011 年 1 月教育部正式批准筹建，标志着全国第一个教育类重点实验室的建设工作正式启动。2016 年 12 月 23 日，实验室顺利通过了教育部科技司专家组的验收。

（二）建设规划

实验室由信息学院、教管学院、高等教育区域发展研究院等学院共建的。实验室也是一个协同体。

建设模式：一体两翼，多条腿；开放合作，谋共赢。

建设思路：以社会需求为导向、协同创新为路径、课题项目为纽带、特色发展为重点、成果实效为目标。

队伍构成：多学科融合，校内外并举，专兼职互补。

图 1 所示为重点实验室的整体构架，它以教育部少数民族教育资源共享平台、云南省教育智库、云南省教师教育联盟等作为支撑的平台，通过信息技术、教育、民族等多学科的融合，与省内外共建合作单位，在省内相关的学院构建研究基地。在此基础上建立了相应的成果转化基地，主要是以企业为主，另外，针对民族的自治州、自治县，以德宏州、临沧市、迪庆州为示范基地，这样便构建了支撑平台、支撑学科、技术合作、产品推广的一个构架。

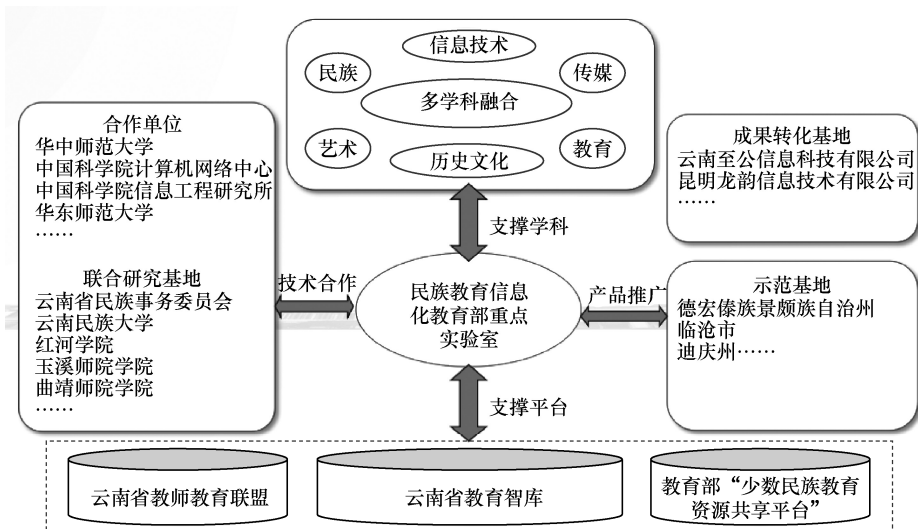


图1 重点实验室整体架构

二、民族教育信息化建设探索与实践

当前，整个民族地区还存在以下几方面的问题：教育基础设施落后、区域教育发展不均衡、教育信息化建设不足。在这种情况下，边疆民族地区的教育仍然落后。

（一）概念界定

民族教育信息化并非简单的民族教育和信息化的结合，也并非教育信息化之前要增加民族限定；民族教育信息化有其独特的内涵，也有其不同于教育信息化的一般规律。

之前，学者们对民族教育的概念观点比较多：民族教育就是少数民族教育、民族教育就是传播民族文化的教育、民族教育就是某一族群传承其传统文化的教育、民族教育就是民族地区的教育、民族教育具有不同层次的内涵等。

通过多年的实践，我们初步对“民族教育信息化”研究与建设有以下四个层次的考虑。

第一层次：民族地区的教育信息化、民族教育的信息化、基础理论及相

关应用研究。

第二层次：少数民族文化的数字化，并应用和推广到各层次教育中。

第三层次：最新教育信息化成果在民族地区的应用与推广。

第四层次：民族地区信息技术人才的培养，特别是基础教育中小学教师的信息技术提升。

因此，民族教育信息化建设是一个整体性、框架性、涵盖面广的工作。

（二）方向定位

主要的工作包括以下 5 个方面：

- （1）民族教育信息化理论研究与实践。
- （2）民族教育信息化应用软件设计与开发。
- （3）少数民族文化保护与传承。
- （4）民族教育信息化人才培养。
- （5）教育信息化的社会服务。

图 2 给出了在民族教育信息化这个大的范畴中，我们规划的建设内容：从基础设施、信息网络、资源建设、相应的软件开发到人才培养的一个大的框架，在此基础上，用信息技术与民族教育进行深度融合，实际上民族教育信息化的根本问题就是“怎么化”。通俗来讲，就是我们如何进行信息化，“如何化”才能使各类、各层次的学习者满意。

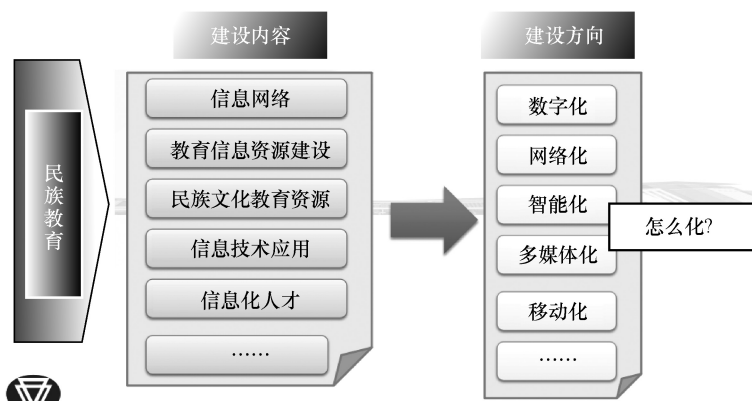


图 2 民族教育信息化建设内容

（三）研究方向

在实验室的建设过程中，我们总结和提炼了以下4个方向。

方向一：民族教育资源数字化研究。

方向二：民族教育信息化学习环境研究。

方向三：信息环境中的学习行为研究。

方向四：民族文化资源数字化应用。

在这个研究中，融合了信息技术、教育、心理、艺术、传媒等相关学科方向，归根结底，目标是深入研究信息技术，推进民族地区教育均衡发展的

问题。

（四）“多元立体信息化”模式

“多元立体信息化”模式即通过政府、高校、社会等多主体的全面参与，依托信息化手段，贯通各级各类教育立体交叉的资源壁垒，建立各级各类教育资源要素自由流动的“教育资源优化配置”机制。

信息化服务边疆民族地区教育均衡发展的探索实践，构建了集“教育信息化基础能力提升、民族地区教育资源开发利用、区域性教育资源优化整合”三位一体的边疆民族地区教育资源优化配置模式。在此模式下，主要的工作是构建“五化五融合”区域性基础教育资源共享机制，探索信息化服务边疆民族地区教育均衡发展的新路径，创新以信息化为基点的边疆民族地区教育师资队伍培训的新方式，厘清信息化服务边疆民族地区教育资源配置的基本理论问题。“五化”主要包括课程学习在线化、教育平台网络化、教学资源数字化、学习方式移动化、网络模式信息化。“五融合”包括创新人才培养与边疆现实需求相融合、教师专业发展与学生能力提升相融合、信息硬件建设与教育资源开发相融合、民族团结示范与教育均衡发展相融合、多元主体联动与产学研政协同相融合。

1. 民族教育信息化理论及应用研究

在民族教育信息化建设的实践中，我们团队主要致力于理论与应用研究，形成了本科（教育技术学、计算机科学与技术）、硕士（教育技术学、计算机应用技术）、博士（教育地理学）为主的人才培养体系。另外，对民族教育信

息化的应用进行了应用软件开发，对昆明市、临沧市、文山州、迪庆州等地提供了信息技术服务。

我们在科学出版社出版了“民族教育信息化文丛”，这一套书是这几年的工作总结和创新，融合了“互联网+”、云计算、教育大数据、移动学习等新技术方法，多视角研究边疆民族地区民族教育信息化建设的发展战略与建设问题。

2. 基于平板电脑的农村学校学生自主学习研究与实践

针对农村小学缺乏专业教师的现实情况，探索教育信息化突破农村教育难点，深入研究与实践基于平板电脑的农村学校学生自主学习模式。该模式应用效果好，共计服务 36 所学校，1288 名学生，为边疆民族地区基础教育师资不足的现状提出了有效的解决方案。该项目受到教育管理部门的肯定，目前在部分县市进行试点，并将进行普遍性推广。学生通过用平板电脑学习，在非专业教师的指导下，信息技术课程、外语课程的教学质量有明显的提高。

3. 开发民族教育信息化应用软件

近几年，我们为云南省教育厅、昆明市教育局及相关的企业部门开发了许多教育信息化的软件，也有很好的应用，其中最主要的是获得了大量教育教学相关的数据，为教育大数据的建设与研究奠定了坚实的基础。

4. 少数民族文化数字化保护与传承

主要运用数字媒体技术、网络技术，实现民族文化信息资源的有效整合，充分发挥民族文化资源传承、利用、开发和共享的作用。开发了大量少数民族数字化资源，这些资源首先通过互联网的共建共享的传播教学，融入到高校、基础教育的中小学课堂中，这将对民族文化的保护与传承起到重要作用。

（五）思考

由于区域性的特点，边疆民族地区在这样的环境下，还有很多工作需要去做，如信息技术与教育教学的深度融合问题、民族地区教育大数据的采集及应用问题、信息化教学环境下民族地区学生心理行为研究等问题需要进一步深入研究。

三、下一步工作思考

(一) 民族教育信息化下一步方向

- (1) 民族教育信息化高水平研究基地。
- (2) 云南省教育信息化智库。
- (3) 建设云南省教育大数据研究院。
- (4) 教育信息化产品研发与成果推广基地。
- (5) “互联网+教育”在民族地区的实践。

(二) 重点工作

(1) 云南省基础教育教师信息技术提升网络学习平台、资源平台、虚拟学习社区建设。

(2) 云南省贫困偏远山区教学点的数字资源服务工程。

(3) 云南省高校状态数据库建设。

(4) 云南师范大学附属小学全面信息化建设，建立基础教育信息化示范基地。

应该说，在“互联网+”下，教育信息化有长足的发展和巨大的机遇，但云南省的区域特点面临着教育质量提升的紧迫性和民族教育改革的艰巨性，在这样的大环境、大背景下，真正的民族教育具有基础性和信息技术的普适性，教育信息化推进云南边疆民族地区教育均衡发展有现实意义，也应该大有可为。我们团队将继续努力，为云南的民族教育信息化事业做出更大的贡献。

作者简介

甘健侯：破格教授、博士、硕士生导师，云南省中青年学术与技术带头人、云南师范大学青年骨干教师；2006年至2008年曾任云南省德宏州陇川县人民政府副县长（挂职），2010年入选中组部“西部之光”人才培养项目，2013年破格提升为教授；现任民族教育信息化教育部重点实验室常务副主任，云南省高校民族教育与文化数字化支撑技术工程研究中心主任。

一直从事教育大数据、民族教育信息化、智能信息处理和数据库技术等方面的研究，主持国家自然科学基金 2 项、国家软科学 1 项，教育部、云南省应用基础研究等省部级项目 10 余项，作为主要负责人承担国家科技支撑计划项目 1 项、国家科技惠民计划项目 1 项。公开发表论文 50 余篇，SCI、EI、ISTP、CSSCI 检索与收录论文 20 余篇，在科学出版社出版《民族教育资源数字化建设与服务》《本体方法及其应用》等专著；在云南省科学技术奖励中获科技进步类三等奖 2 次，自然科学类三等奖 2 次；为国家汉办、云南省教育厅、昆明市教育局等部门开发软件 40 余项，获得国家版权局计算机软件著作权 17 项。

大数据在教育中的应用及限度

曲阜师范大学教授 谭维智

教育大数据主要包括教学过程的数据和育人过程的数据，按照学习过程中的人分为学习者的数据和教育者的数据，按照教育功能分为教学数据、管理数据和评价数据等。对于每一种数据都能进行细分，如学习者的相关数据可以分为学生基本信息（座位信息、学习经历、认知状态等）、学生行为数据（参与学习社区、在线互动情况记录及应答情况）、学习环境数据（课程目标、评价指标、学习资源、课程资源等）。

目前，教育大数据应用最大的问题还是缺乏有价值的信息，如教学过程的数据，教学数据的采集主要还是依赖在线教学的普及。如今，在中小学和大学在线教学的应用还是很有限的，只有真正用起来才能产生数据，我们才能采集到有价值的信息。大数据在教育中的应用具有非常重要的价值，基于大数据的个性化教学、科学化的评价、学校的精细化管理、智能化决策及精准化的教育科研等，对于促进教育公平、提高教育质量、培养创新人才都具有重要意义，宏观层面还可以为教育决策提供支持，提供监控、管理，优化教学过程；微观层面可以提供个性化的学习诊断和各种干预的策略。目前教育大数据的应用还包括教育舆情的监测、教育热点问题的追踪、国民体质的监测等，这些都是中国教育大数据研究院正在进行的大数据研究项目。

教育大数据主要建立在学习分析的相关平台架构上，否则，便无法进行系统的挖掘，技术的瓶颈的影响还是比较大的。目前应用于教育采集的技术主要包括以下几类：第一类是物联感知技术，如校园一卡通，这种一卡通可以积累很多有用的数据，可以利用它分析学生在校内活动的轨迹、借阅图书的数量、在校园餐厅消费刷卡的情况等；第二类是视频技术，这类技术目前运用得越来越普遍，很多学校的教室、活动场所都装了监控，可以监控每一个班级学生上课、教师授课的情况，录制的视频除了可以存下来观看，还可

以结合人工智能技术对视频进行分析，如通过人脸识别技术可以分析学生是否在学习、其注意力是否集中；第三类是图像识别技术，如采用数码笔将学生的作业上传到云平台，接着在平台上进行相关的分析；第四类是平台采集技术，通过建立各种平台及日志记录的方式，记录学生在平台中的一举一动。除此之外，还有一些技术对教育有着非常重要的意义，如人的情绪数据采集技术，这种技术主要是综合人的脸部动作来判断一个人的情绪，这类技术可以涉及人的意识领域，在教育领域的应用前景是非常广阔的，甚至比一般的学习分析更有意义、更有价值。

通过教育大数据技术，可以量化过去无法量化的信息，使用精妙的统计学分析方法分析教育过程中的各种信息，而过去对教育过程中的各种信息，主要是靠经验、靠估量、靠课堂上学生举手和大体比例，例如，语文老师布置学生在课下预习课文，但学生具体的实施情况是无法得知的。再如，课上学生对教师授课的理解程度基本是靠估量的，教师很难精确地掌握学生的理解程度。学生上课的听讲情况、注意力是否集中，只能靠观察来掌握大体情况。传统的教学中，由于班级学生人数众多，教师凭经验、传统的手段很难做到因材施教，一个教师面对七八十个学生，教师个人的能力和经验难以做到因材施教。尽管教师可以通过分析学生成绩来了解学生的学习情况，但也只能帮助教师初步了解大多数学生的共性问题，很难精准地掌握每一位学生的情况。随着数据存储技术、自然语言处理技术、语义分析技术、数据挖掘技术等大数据技术的不断发展，利用学习分析技术可以全面记录学生的学习行为，可以利用这些数据分析学习者的学习习惯、学习方法及潜在问题，运用这些分析的结果为学生提供个性化的学习内容和自适应的学习方式。在全媒体教室中还可以实时监测学生的注意力情况，通过音频、视频数据的分析记录下学生的整个学习过程，对其学习过程进行分析。未来人工智能还能对学生的表情、语气中的情感进行分析，可以进一步掌握或研究分析学生在学习过程中的思维转换与意识。2016年，哈佛大学做了一个注意力检测头环，通过检测脑电波，并和教师的电脑联网，教师便可以及时注意到学生上课的注意力集中情况，这在平时的课堂上是几乎不可能实现的。

目前，我们对教育大数据的应用还处于初级阶段，大多数学校只能提供教师备课数据库、考试数据库，稍微好一点的开始采集学生学习过程的数据，

如通过扫描仪扫描学生的试卷，在机器上阅卷并进行分析。再如，采集平时学生练习的数据，把这些数据集中起来便可以进行分析。

从目前大数据在教育过程中的应用来看，对宏观层面把握确信的教育教学状况，了解和分析大范围教育教学趋向共性问题具有非常积极的作用。因此，目前我们对大数据分析得出的结果，仅仅是一般性情况，考察的是大势，分析的是一般性规律。但教育的应用不能仅满足于掌握大势，我们要注意到教育的对象是人，最终要落实到每个学生身上，教育中的每个人都是不同的，每个学生都是完全不同的个体，只有符合学生特性和兴趣的教育才是有吸引力的。因此，教育需要研究大数据，也需要研究小数据，大数据找规律，小数据找痛点，只有找到学生的痛点，才能对症下药，解决学生的问题。必须注意到教育具有特殊性，教育的对象是人，人是具有想象力、创造力的动物，每个学生都是不一样的，而且课堂上发生的事情都是偶然性的。因此，除了关注教育大数据之外，还要关注数据的研究。

2011 年，乔布斯在与比尔·盖茨的会面中讨论了关于教育的问题和未来学的设想，他们一直都认为计算机对学校的影响小得令人吃惊，这是相对于媒体、医药和法律等其他领域而言的。美国联邦教育部部长邓肯也提出了一个类似的问题，即为什么各国政府教育信息化投入了很多，但产出和投入却不成比例。事实上，直到今天，我们在教育技术领域也无法拿出令人信服的、具有革命性的信息技术与教学能完美融合，还没有一个研究能证明学习成绩的提高确实是应用计算机的结果。

科技的发展、条件的改善，使人的体能、智能退化，会使人变得更懒惰，这是无须证明的问题。例如，空调的使用会使人的抗寒、抗暑能力下降。电脑的发展是否会使人智能下降是值得我们注意的问题。2017 年 6 月，苹果首席执行官库克提出这么一个问题：“我不担心人工智能会让计算机像人一样思考，我更担心人类像计算机一样思考，失去了价值观和同情心，罔顾后果，这是我需要大家帮助预防的。”就教育领域对计算机、互联网、多媒体等技术的使用而言，其产生的后果目前还不得而知，多媒体对学生而言也不一定是隐患，比如数学课学生需要进行很多演算，教师需要挖掘隐藏在定理背后的思想，这并不是 PPT 能做到的，PPT 仅仅是教学辅助书，永远无法取代传统的教学模式，它能发挥的作用最多是将概念、定理及静态和动态的图片事先

放到屏幕上，以便于节省时间，也可以完成传统板书完成不了的工作。但现实是多媒体不仅取代了板书，也限制了教师的思想，学生在课堂上看到的是与教材一样冰冷的符号与文字。

另外，数据分析只能告诉我们“是什么”，却不能告诉我们“为什么”，在教育上“为什么”要比“是什么”重要得多。学习的意义也不限于知识，但目前大数据的学习分析更多是对显性知识及学生掌握这些知识的行为过程进行分析，它在给学生和教师提供精准的学情教程的同时，也将经常性地灌输教学导向。就人的学习而言，学习的意义几乎等同于生活，人的学习很多时候并非发生在学校、在使用计算机设备的时候、在教师教的过程中，学习中最重要东西无法被数据采集到，只有在使用计算机、智能设备、互联网的时空才能产生我们需要的大数据。教育和学习的时空显然是互联网和现有的智能设备无法完全覆盖的。

数据是信息和知识经过表达、一定的加工或编码所形成的，能进行数据处理的都是可以表达出来的东西，但在教育上有很多是不可教、不可说的东西，也无法进行数据的处理，如作业的批改、作文的批改等都是技术的难点。在教育过程中，除了我们所注意到的之外，还有很多偶然的学习，在学生的一生中，偶然学习的成果可能比正式教授的主要材料还重要。例如，在学习历史时，一个学生可能附带学习到了有效的学习习惯、对学校的良好态度或对整个学习的爱好。但经过学习，他可能考虑到为了争取好分数可以不择手段，学生大量的、偶然的学习来源于看不见的课程，这是无法量化的、数据也无法采集的东西，如教学大纲或课程进度表并未明文规定的、教师不讲述的甚至不知道学生正在学习的东西等，所有这些都是教师在毫无察觉的情况下进入了正式的课程，这样便导致学生看不见，教师也看不见，大数据的研究也采集不到。

通过现有的技术手段能采集到的学习数据是非常有限的，基于我们对学习的认识、技术使用的范围，那些偶然发生的学习都是很难进行数据采集的，同时，这些也是非常重要的。我们目前可以采集到的数据仅仅是各种数据中的冰山一角。值得注意的是，有时我们为了采集数据不得不采用计算机、网络等各种电子设备，课堂教学中要使用网络、多媒体教学手段，这些手段在帮助我们采集到数据的同时，其效果是不得而知的。例如，国际金融组织关

于电脑使用对于学生的影响相关研究发现，越是提升学校的学生与配置计算机台数比例的国家，学生成绩越呈现下降的趋势，其中也出现学生使用计算机频率越高，阅读能力越低的情况。

很多研究证实，互联网妨碍了学习，网络将超文本技术与多媒体技术融为一体，用来发送所谓的超媒体内容，超媒体不仅是以电子形式连接起来的文字，还包括图像、声音、视频等。很多教育专家认为，多媒体会加深理解的程度，强化学习效果，输入越多，效果越好。但也有很多研究得出了与此相反的结论，如有些研究已经证实，多媒体所要求的精力分散进一步加剧了认知的疲劳，从而削弱了学生的学习能力，降低了学生理解的能力。所以，我们为大脑提供的思考原料，并非越多越好。

大数据在教育中的应用不同于其他行业、其他领域的应用，培养人的方式与制造业、通信业、交通业有本质的不同，这些行业的对象本身就是技术的产物，甚至是技术本身，作为技术的指向对象可以被动地被改造，而教育的对象指向的是人，人不是技术的产物，更不是技术本身，人具有主观能动性，人只能自我改变、自我改造，人不可能被技术改变。人的学习本质上只能是自我练习、自我学习、自我成长，机器和技术不能取代人的自我学习，其他人也不能替代学生的学习和成长，人只能通过自己改变自己，而不能被技术改变。

当今很多一流的科学家告诉我们，大脑最深层的奥秘是非计算的，由于大脑意识活动的不可测，实际上也限制了我们获取人脑学习相关数据的可能性，我们能够获取的都是低质量的、边缘性的、间接性的数据。以心理学的研究为例，得到的数据都是大脑某个位置的电波，距离记录人脑真实的思想内容还很远，大脑对我们而言还是一个黑洞，人类的思想意识在其中的流动运行，我们几乎一无所知。而这正是教育至关重要的层面，是教育的根本。教育不仅仅是知识的堆积和积累，这不是人存在的根本，这些任务目前完全可以交给电脑和互联网来做，教育根本性的任务 and 目标是培养人具有正确的价值观和道德品性，培养人心中最核心的判断力、创造力、自由的精神、独立的人格，培养人的良知和良性，而这些对技术而言还无能为力，属于技术一无所知的区域。技术背后包含知识，技术也可以操作知识、传递知识，但技术不能操作价值观，不能通达人性，技术只涉及了人性中最微不足道的角

落。多媒体技术、信息技术、人工智能等对教育而言仅仅为一种工具，其价值只体现为价值工具，是一种学习的助长手段，但教育本质上是人的自然教育，学习本质上是人的自我学习，在教育和学习的问题上恰恰是一切其他人和外在的工具手段无法助长的。

大数据的应用将使我们更加清晰地认识教育的本质，在大数据时代、技术无孔不入的情况下，教育的目的可能是保持人性，特别是保持人之为人的天性、天赋、好奇心、求知欲、想象力等，这些是人区别于机器，不能被机器取代的根本所在。

作者简介

谭维智：曲阜师范大学教授、博士生导师，中国教育大数据研究院首席专家。山东省有突出贡献中青年专家。

近5年以来，主持国家社科基金项目2项，教育部人文社科项目1项，横向课题1项（到账经费20万元），参与省部级课题4项。出版著作4部，其中专著1部（独立作者）、合著1部（第一作者）、参著2部。在《教育研究》发表论文5篇。这些成果在学术界和社会上产生了较大影响，被《新华文摘》全文转载3篇、观点摘编1篇，人大复印资料转载5篇。研究成果获国家级奖3项，省部级奖3项，厅级奖5项，分别为山东省社科优秀成果奖一等奖2项、二等奖1项，全国优秀博士论文提名论文1项，全国教育科学研究优秀成果奖一等奖1项（位列第6），山东省高校优秀科研成果奖一等奖3项、二等奖1项、三等奖1项。

大数据背景下的“新工科”培养模式 ——以软件工程为例

曲阜师范大学软件学院教授 倪建成

曲阜师范大学始建于 1955 年，由周恩来总理 1954 年访问曲阜时亲自规划并选址。曲阜古为鲁国国都、孔子故里、儒家文化发源地，现为世界遗产城、特色魅力城和国家历史文化名城。

自 2006 年曲阜师范大学设立软件工程专业，尤其是 2014 年成立软件学院以来，专业建设紧跟产业需求和软件工程教育发展潮流，力求在培养目标、知识体系和培养模式等方面与国际接轨，形成了颇具特色的软件工程应用型人才培养模式。

以新技术、新业态、新产业和新模式为特征的新经济是发展新动能的源泉，是抢占未来产业、工业和科技革命制高点的关键要素。人才作为发展壮大新经济的首要资源，既需具备工程学科的专业知识，又需具备工程思维和工程实践能力。

然而，在我国已拥有世界最大规模工程教育的背景下，迅猛发展的大数据、物联网、人工智能、网络安全等新经济领域都出现国际化、工程化人才供给严重不足的现象，暴露出我国工程教育与新兴产业、新经济发展脱节的短板。造成这种局面的主要原因在于专业培养模式重视学科导向而忽略产业导向，遵从于专业分割而忽视跨界融合，面向被动适应而非引领支撑。

因而，为满足新经济形态下的国际化、工程化、职业化人才需求，工程教育必须主动适应，面向产业需求深化教学内容、课程体系、实践体系和科研体系改革，以学科前沿、产业和技术最新发展推动教学内容更新，把职场环境引入教育全过程，把创新创业教育融入工程教育全过程，强化工学结合，充分发挥双师型、多学科师资优势，基于能力对学科知识的需求，构建柔性、

灵活的知识体系，改变学生能力评价的方法和考核体系，培养学生在个性化学习中发现科学问题并能运用科学原理解决问题的终身学习能力和知识传播、共享能力。

此外，我国校企政合作历史虽然已有二十余年，但高校目前面向社会汇聚优质资源的动力和能力依然不足。为推进工程教育培养模式改革，高校应在建立开放标准的基础上，推进校企政一体化合作，充分发挥工程教育在师资队伍、实践平台、行业协同等方面的优势，在更大范围内优化配置教学、实践资源，以内外资源创条件，打造工程教育开放融合新生态；高校应与行业企业、科研院所合作，以产业发展和社会需求为导向，探索建立产学研合作协同育人的长效机制。

我们深知信息产业发展迅速、日新月异，只靠经验判断信息产业的人才能力需求和技术需求，判定信息技术发展趋势，难免会出现培养模式僵化、自适应调整匮乏的窘境。随着近年来大数据技术的快速发展和应用，我们积极调整思路，一直致力于尝试将大数据技术应用与软件工程学科和专业建设相结合，这也与“新工科”理念不谋而合。

在笔者 2015 年提出的“职业岗位能力需求反推课程体系”的弹性培养模式基础上，2017 年又进一步提出了面向新经济的“二多三化”培养模式。“二多”指多学科、多合作主体；“三化”指国际化、工程化和职业化。设计开发了以软件工程学科为例的综合性、普适性工程型人才教育云资源平台。

通过构建面向职业岗位能力需求的学科、项目交织的柔性、模块化的弹性工程教育课程体系，既有助于改革教学方法，又有助于满足新经济背景下的跨学科人才需求；通过搭建校企政深度协同的云资源平台，既能满足工程职业实践为基础的“新工科”人才培养内在要求，又能满足校、企、政、生的利益诉求。从而能动态适应现代工程企业人力资本需求日益多元化、个性化的发展态势。

下面重点阐释云资源平台。软件工程教育云架构主要有 4 个相互协同的子系统，分别是学科舆情监测系统、知识自适应拆组系统、教学过程系统和能力认证系统。

学科舆情监测系统是教育云平台的“侦察兵”，它充分利用了互联网不受时空限制、信息更新速度快、资源整合程度高等优点，实时监测企业、行业、

社会对学科人才培养规格需求，实现收集职业岗位生态及其环境信息的目的。针对软件工程学科人才，舆情监测系统的检测范围不仅包括智联招聘、前程无忧、猎聘网、BOSS 直聘等专业招聘网站，而且包含 BAT、京东等一线互联网企业的招聘公告，通过分析允许使用范围内的数据，用于掌握职业岗位的最新能力需求信息。

同时，舆情监测系统具有自校功能，能结合系统历史信息，利用数据可视化技术刻画出同类岗位能力需求的变化趋势，自动生成近年来岗位能力需求报告；通过分析学科知识、技术热度，预测专业知识在未来一段时间的发展趋势，从而不再闭门造车，让培养方案、课程体系的规划工作有据可依。最后，舆情系统自动生成学科、专业对应的职业岗位能力需求报告，并自动传递给知识自适应拆组系统。

知识自适应拆组系统依据职业岗位能力需求报告，将课程体系拆分成课程模块-能力培养矩阵。课程模块包括通识模块、学科核心模块、专业拓展模块和工程实践模块；能力主要包括学术、实践、专业综合等技术能力和国际视野、创新、团队协同等非技术能力。

继而，结合学科教育知识体系，系统进入课程模块的知识域、知识单元、知识点的拆组阶段。对软件工程学科而言，教育知识体系主要来源于 CBOK、SWEBOK V3 版等规范。值得一提的是，为了达成多学科深度交叉融合目标，体现学校的办学特色，满足特定应用领域和产业需求，课程体系引入了行业和企业技术，设立了满足行业发展的课程和项目实践模块，从而既实现了知识体系的跨界交叉融合，又满足了不同能力结构的培养需求。

教学过程系统接收到知识自适应拆组系统的课程知识点报告后，进入教、学阶段，教育云资源平台提供了强大的以知识点为最小单位、对学生进行系统工程教育的支持服务。教育云教学过程采用 5 层微循环递增策略，分别是电子教案预习、模拟课堂教学、在线讨论、知识水平测试和知识回归反馈。此外，实施教、学过程的全程监测，真正做到了教、学状态的实时跟踪和溯源。

教学过程系统内置了 7 种学习模式，包括被动统计学习、主动学习、算法式教学、演示学习、感知因果学习、因果学习和增强学习。值得关注的是，其中的算法式教学、演示学习、因果学习和增强学习是充分利用了教育云平台特性与优势，为工程学科量身打造的学习模式。

比如，算法式教学是指系统基于职业能力需求目标，在分析学习进展情况和已拥有能力水平的基础上，合理规划后续课程知识点并设计案例，指引学生进一步达成个性化的学习目标。

演示学习是指将难以理解或难以用语言描述的工程实验，以视频、图像、代码等直观展现方式复现操作过程，从而降低学习曲线复杂度。

此外，教学过程系统还广泛采用了数据可视化技术。数据可视化技术的直观化、关联化、艺术化和交互性的特点，为处理教学过程中的数据提供了良好的支持。

数据可视化技术可作为一种认知工具，用以支持、展示、指导和扩展学生的系统化思维水平。例如，基于学生原始的学习记录数据，通过分析、过滤、挖掘、表述、修饰和交互等操作，系统可以为学生直观地呈现学习进度和目前的学习态势，作为学生自我评估的参照，使学生及时发现学习过程存在的问题，了解自己的学习行为，如学习时间、浏览内容、学习质量等，并与其他同学形成比较，反思自己的学习过程，促进学生自我规划，成为主动的学习者。

将经过拆组后的知识点生成知识点体系图谱。每个知识点的掌握程度都可以热力图的方式显示在知识体系图谱上，知识图谱的颜色分布和深浅程度，生动、形象地展现了学生对课程的掌握情况。当然，学生也可以通过对探究性数据的可视化，加深对知识的理解，完成知识的显性直观化，增加知识的交互性与关联性。

数据可视化也可以让老师实时根据学生的实际情况进行课堂教学、教学干预和教学评价等。教师可以为学生呈现直观化的知识，加深学生对于知识的掌握程度；可以通过对学生学习数据的分析，为学生提供个性化的指导；还可以将其作为一种评价学生的手段，进行总结性评价和过程性评价。

教学管理者还可以掌握教师教学效果和学生学习情况，人工干预教学管理与决策的目标、方法和策略。单个体行为数据似乎是杂乱无章的，但当数据累积到一定程度时，群体的行为就会在数据上呈现出一种秩序和规律。分析这种秩序和规律，有助于提高管理者决策的准确性。

能力认证系统借助教育云强大的数据存储和计算能力，根据学生用户在线学习记录的全过程数据，如在线时长、鼠标键盘操作、在线提问质量、知

识点掌握程度和知识点测验结果等指标，客观分析学生学习的实际情况，计算学生的能力范围和水平，科学评价学生的职业岗位能力，从而既可以对学生学业进行预警，也可以向学生推荐与之能力最匹配的岗位列表。

对于能力较弱的学生，可重新进行知识点拆组服务，系统将针对学生能力薄弱的知识点，重新组合成为新的课程模块，供学生继续在线学习，提高职业岗位能力。知识点自适应拆组服务也为学生的终身学习提供了一种可行范式。

目前，能力认证系统正在进行积极探索培养应用型人才的金字塔证书体系，按照学生职业能力水平颁发相应资质证书，建设类似于 Oracle 认证、华为认证、微软认证的综合能力认证平台，为向企业和社会提供对口工程型人才提供可靠的质量保证。

综上所述，我们将大数据背景下建设的“新工科”教育云平台的优点归纳为以下 6 点。

第一，灵活的课程体系，弹性的培养方案，科学的论证体系，完善的舆论监督，合理的风险规避。

第二，强大的知识自适应拆组功能，为每位学生量身打造个性化的成才之路，避免一刀切、填鸭式教学。

第三，多态教学模式，全方位教学手段，全过程学习监督策略，实时掌握学生学习情况，达成能力提升目标。

第四，严谨科学的能力认证体系，弹性灵活的学习机制，与时俱进的终身学习模式，确保每位毕业生的培养质量，对学校培养工作负责、对学生能力提升负责、对企业人才满意度负责。

第五，彻底的“新工科”培养模式，按需实施的课程体系，及时反映技术发展新方向、企业新要求，实现校企无缝衔接。

第六，校企资源共享，校企深度融合，达到校企生多主体共赢目的。

曲阜师范大学历来十分重视教学改革，尤其在国家提出“互联网+”“大数据”“新工科”等重大发展策略后，软件学院成立了包含企业高工、软件行业专家和优秀专业教师的教学改革研究团队，学校投入巨资建设包括云计算中心在内的基础设施之外，一直给予软件学院“特区”政策。相信在国家“新工科”理念的指引下，我们必能“学而不厌，诲人不倦”“不忘初心，继续前行”。

作者简介

倪建成：博士，曲阜师范大学软件学院教授，中国计算机学会高级会员，IEEE 会员。主持省研究生教育创新计划项目和横向课题各 1 项，参与国家自然科学基金项目和省厅级项目 10 余项；发表论文 30 余篇；获省厅级奖励 6 项。主要研究方向：分布式计算、信任计算、数据管理与决策。



人 才 篇

大数据人才培养之道



“互联网+”时代创新人才培养模式的思考

佛山科学技术学院校长 郝志峰

如今，我们面临的一个严峻的局面是：“互联网+”时代的原住民究竟是谁？当我们谈论社会主义事业的建设者和接班人时，往往谈论的是互联网时代的社会主义。教师和学生谁会是“互联网+”时代的原住民？乔布斯曾有一个著名的问题：为什么在教育领域的信息投入很大，但却未产生类似于生产流通领域的效果？其中的核心问题——教育是滞后的。

一、广东省“互联网+”时代教育改革实践

当我们讨论到“互联网+”时代创新人才的定义时，党和国家领导人提到的两大技术是不容忽视的，即非对称技术和硬技术（或颠覆性技术）。目前我们一直思考的是什么是“双一流”、高水平下的人才培养。如今，著名大学和一般性的大学都在谈双一流，想要处理好的基础与应用、冷门与热门、长线与短线、自然科学与人文科学一直是我们长久以来不可回避的矛盾。因此，当我们考虑和犹豫拔尖创新人才时一直处于徘徊期。围绕“双一流”国家分了多个层次，如典型的 985 高校中的 C9 联盟也在谈论“双一流”。2016 年“两会”期间，广东省教育厅面向全国首先发布了高水平大学和高水平理工科大学建设的广东经验，有两所教育部和五所地方的共七所高校。佛山科学技术学院作为一个地方高校，如何进行高水平的大学建设或高水平的学科建设？需要冷静思考。因此，在整个佛山科学技术学院建设高水平理工大学的过程中，最重要的是为佛山服务，为佛山的制造业大市服务。

广东在整个高水平理工大学建设的工程中也有很多新的提法，如应用型高校的转型。2016 年 12 月 20 日开展了 12 所投入 100 亿元的共建本科院校，即在每个地级市都有一到两所合作的本科院校。在 2016 年 12 月 31 日，广东

省的高校又新增了一批重点学科，广东省的重点学科面几乎达到了全覆盖。因此，从2014年开始，广东省已经推出了“创新强校”的理念，其中的“创新”既有科技创新，又有人才培养创新，或是创新人才的培养。

目前，教育部积极推动200所高校向着应用型高校转型，应用型高校在转型过程中首先要思考几个关键词，即新产业、新业态、新技术。换言之，大数据、互联网+、移动互联网、物联网，包括人工智能、云计算等是否属于新产业、新技术。另外，应用型高校能否紧密对接产业链、创新链也是说易行难的，高校本身并不处于产业链和创新链的顶端，如地方高校、985高校确实会存在这样一些思考。因此，教育部在引导部分地方普通本科高校的应用型转变的指导意见中，提到了“互联网+”时代培养创新人才的困难之处：第一，应用为驱动，即培养应用型、技术型、技能型人才；第二，围绕人才培养的方案和课程进行改革，加强实验和实训机制建设，包括双师型队伍的建设。因此，广东省教育厅从2016年暑期开始也根据教育部文件出台了广东省实施意见，最终将14所高校选为转型高校。从学校分类角度来看，至少可以分为以下几类：世界一流的大学或世界一流的学科、高水平的大学或高水平的学科、地方的、国家的、国际的，以及所谓的创业型大学、层次性大学，还包括职业教育培养的立交桥，有高职生、本科生、硕士、博士等。

二、创新人才培养模式的探索

（一）教育部关于人才培养的思考

提及创新人才每个人都很渴望，但对创新人才培养模式的探索会变得极其困难。一个核心问题是创新人才是谁培养的。目前国内外的人才大部分是通过学校培养的。那么，学校在人才培养的过程中出了哪些力？教育部做了如下几方面的思考：①自主学习的模式。换言之，千学万学、千教万教不如学生自学。②诸多教学方法的改革。如启发式、探究式、讨论式、参与式。③考试方法的改革。教学改革是否花费了足够精力研究考试方法的改革，尤其是注重学习过程的考察和学生能力的评价，我们都会幻想考试分数能与学生的能力间表现出某种大数据的统计关联关系。就事实而言，长久以来考试效率很高，但缺乏的是对学习过程的考察。

（二）人才培养模式的核心要素

围绕着“中国制造 2025”“工业 4.0”时代，教学中的“教”与“学”应分开，过去的教学改革大部分是“教”的改革，如今进入“互联网+”时代、大数据时代，才开始有了对“学”的过程把握，当谈到“工业 4.0”人力变为机器的时代，教学是否也从人力变为机器？换言之，教学究竟是 1.0，还是 2.0？因此，不论是直接面向“中国制造 2025”的机械类专业，还是黄淮学院牵头的产教融合的国际论坛，一直都在思考同样的问题。而人才培养的模式（“教”与“学”的过程）是按照德国以柏林大学开创的现代大学体系——专业、课程与课堂 3 个核心。那么，“互联网+”时代面临大数据环境，我们对专业课程和课堂的改革有何认识？目前所有大学的基本特点都是专业毕业，不论是大数据专业，还是数据科学专业，仅仅是学了一个专业，而一个专业仅仅是全校几千门课中的一个子集和，解决这些课程的困难和矛盾是每天与学生面对面的课堂，因此，人才培养模式突破的核心要素仅有 3 个内容：专业建设、课程改革、课堂实践。

在专业建设之中：“钱学森之问”是不可回避的事实，中国没有一所大学、一个大师能满足其创新人才培养理念，但他在中科大的实践给我们带来一个启示，即专业的核心问题是各类课程的比例，对于大数据专业，我们能否分清数据科学、大数据等各类课程的比例，以及统计计算机、人工智能、机器学习等课程各占的比例，甚至包括数学基础课。

在课程改革方面：“互联网+”时代背景下除传统意义上熟悉的理论课，实验课、理论实验并重课、实验算法学都是值得思考的问题。

在课堂实践方面：课堂更是目前我们所面临的最严峻的考验。不论是学生为老师打分，还是教务处的督导都十分关注，除课堂教学效果外，还有作业和教材问题，如 985 高校、211 高校的作业是谁在批改，一般高校的学生作业对其学习是否有帮助。

在“互联网+”时代还未到来时，谈论这些问题时都曾幻想孔子的“因材施教”，但仅仅是幻想。只有在“互联网+”、大数据的时代，“因材施教”才露出曙光。因此，教学，尤其是学习过程才刚刚进入 1.0，还未进入 2.0，因为学习要比下围棋复杂得多。

以美国为例，美国关于高等教育的研究做了一个报告，其中有很多对教育改革有深远意义的观点，如提出设置基础知识的核心课程群，广泛地重视应用信息技术的研究。目前看来，全世界都在考虑下一代学习的挑战，数据科学带来的不仅仅是专业课程与课堂的困难，还有下一代学习的挑战。因此，美国人围绕 STEM 专业课程与课堂建设也在进行思考。1957 年苏联卫星上天，美国开始重视 STEM 课程，1986 年中国国门打开，美国再次振奋。美国认为 STEM 课程是一个很大的难题，关键环节来自于扩招。美国人在思考 STEM 教学时，考虑了一个最重要的突破，即如何使学生在 STEM 课堂中提高设计能力、批判性思维能力、问题解决能力。美国人的解决方案很简单，即 STEAM，在过去的 60 年中，更多地集中在传统的 STEM，老师在课堂上通过一个案例或问题让学生组成若干探究小组，从收集、分析数据到提出解决方案，再到协同学习。美国人认为 STEM 素养不是单靠一门课形成的，而是在分析案例、解决问题的过程中形成的。近期，美国才将 A（art）加入。

传统的 STEM 教学或数据科学专业人才培养方案，要有核心课程，要有能实践这些专业课程的课堂。在“十二五”期间，尤其是在过去的精品咨询课、资源共享课等的推动下，专业课程和课堂改革增加了核心课程群，这与中小学谈论的核心素养类似，而关键难题在于实现专业核心课程群的课程和课堂，这便需要高度关注信息技术，通过信息技术才有可能实现对上述专业核心课程群课程和课堂学习过程的把握。因此，“互联网+”时代数据科学专业如何利用大数据推动教学改革是一个艰巨的难题。

三、“互联网+”时代教学改革的后事之师

以慕课为例，经过 2015 年到 2016 年的“互联网+教育”的鼎峰，如今逐渐回归平静。可以说，目前“互联网+教育”应回归教育本身，“互联网+”仍是工具。2016 年，清华大学颁奖会上讨论的几个议题是“互联网+”时代教改必须关注的几个方面：第一，“互联网+”时代会不会重构教育模式，“互联网+”时代，中美两国围绕专业和课程的建设是否会有新的变化。第二，“互联网+”对教师职业产生的冲击。第三，在“互联网+”时代，教育与人性的融合与博弈，包括一些互联网大佬对人工智能的恐惧感。传统意义上认为慕课的形式

是一种碎片化、离散化的学习方式，但所有的数学课、计算机课、统计课都有相应知识的连续性和集成性。因此，在“互联网+”时代、大数据时代，互联网的思维显得尤为重要。美国宣称，通过慕课，美国的高校可以在全球搜索最优秀的人才，这也是大数据的应用之一。因此，教育部围绕的相关课程中，包括最近推出的精品资源共享课，如数学建模课，均引起了高校的高度重视。

四、结论

“互联网+”时代的教学改革主要有 6 个核心环节。

(1) **专业群**。数据科学专业目前至少有 3 个课程群思路：第一个放在数学中，是北大鄂维南院士提出的；第二个放在统计中，是中国人民大学、上海财大的思路；第三个放在计算机中，这也是传统的数据挖掘课程。因此，当我们谈论“互联网+”时代的数据科学专业时，更多地是谈论其属于哪一个专业群，这也是一种趋势。另外，在“十三五”期间，应高度重视慕课，使数据科学相关课程得到很好的推广。除专业课程与课堂的关注外，还应关注专业群的趋势，关注核心课程群的思想方法，关注慕课的应用和推广。

(2) **个性化学习变成可能**。在“互联网+”时代一个最重要的环节是我们可以面向学习、面向学习过程做一些思考。如目前发现个性化学习变为了可能，以前的课堂基本不允许学生犯错。因此，因材施教，当下才初显可能性。

(3) **师生合作**。老师不仅仅是知识的传授者，还是知识学习的帮助者，同时，也应以自主学习为主。

(4) **学习过程**。以前探索着作业在学生间互改，如今，对互动评分、打分系统、学习系统、评估系统特别重视。尤其是要注意每一位学生的学习效率。

(5) **进行分层次、分班教学**。过去的教学改革流行分层次教学，目前三五个人也可以分层次，对班级的学习台阶进行细分化。

(6) **对学生的学习效率给出针对性指导意见**。这便是“互联网+”时代的教学改革。

作者简介

郝志峰：校长、数学教授、博士生导师、党委副书记。广东省“千百十”工程省级人选、美国数学会会员、教育部 2013—2017 年高等学校数学与统计学教学指导分委员会副主任委员、广东省数学会副理事长、广州工业与应用数学会理事长、中国决策科学研究会理事。获国家政府特殊津贴。广东省超级计算机应用产业联盟的理事长、广东省复杂过程信息物理融合系统工程技术开发研究中心主任，从事高性能计算和海量数据挖掘研究。

大数据应用创新及人才培养探讨

河南城建学院计算机与数据学院院长 何宗耀

河南城建学院作为全国仅有的两所“城建”类院校，在过去的15年中，积极推进转业建设、学科建设、学生培养等，取得了一定的成绩。计算机与数据科学学院在学校转型发展过程中，紧紧结合地方、住建行业的需求和学校转型发展的要求，为培养更多适应“大云物移智”快速发展需求、服务于建设行业要求的应用型人才方面进行了大量的探索。

一、“百校”工程

“百校”工程是教育部学校规划建设发展中心与中科曙光发起的产教融合创新项目，主要是在全国一百个城市，每个城市一所学校，每所学校建设一个大数据应用创新中心，目的是共建大数据协同创新网络。

（一）数据中心服务面向

河南城建学院大数据中心立足于大数据智慧城市建设，面向河南建设领域，目标是服务于智慧城市建设，服务于科研人员学术研究的数字支撑，同时为社会公众对数据需求的关注提供服务。

（二）基础建设

目前，大数据中心已进入设备的安装调试阶段，大数据学院已启动建设。从2016年4月至今，河南城建学院主要做的工作包括大数据中心的基础环境及设备招标、人才培养（大数据学院的启动及数据科学与大数据专业的建设问题）、相关的行业数据源与基础数据源支撑。

（三）数据源建设

数据源的建设主要从以下 4 个方面考虑：网络数据（包括网络的公报、年报统计数据）、行业的统计数据（包括房产、地产、规划、交通、环境、地理信息基础数据等）、官方公布的相关宏观数据（包括国家统计局、河南统计局，以及地方相关纪委公布的数据）、物联网采集的一些社会相关数据。

二、明确方向 准确定位

（一）学院的发展定位

依托学校的建设行业特色，以培养满足建设行业和智慧城市建设需求的具有大数据思维和工程实践能力的应用技术型人才为目标，建设具有行业特色的“计算机与数据科学学院”。

（二）优化专业结构

构建合理的专业结构，按照“新工科”和“工程教育专业认证”的要求优化人才培养方案，专业结构如图 1 所示。

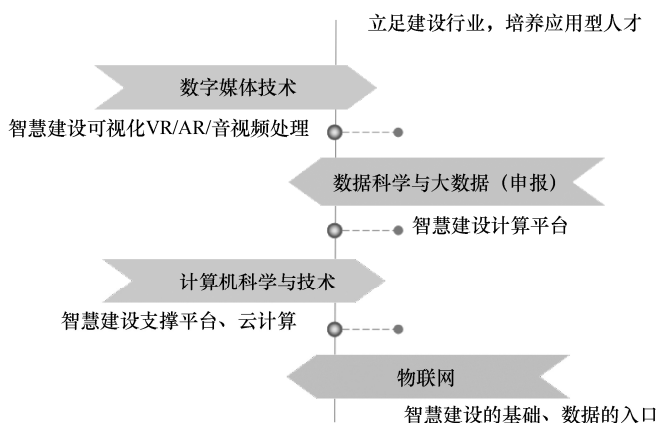


图 1 专业结构

（三）校企行政立体交叉 产学研用协同育人

以校企合作为基础、产教融合为目标，积极推动协同育人、协同创新，

已经建设到位的协同育人平台如图 2 所示。

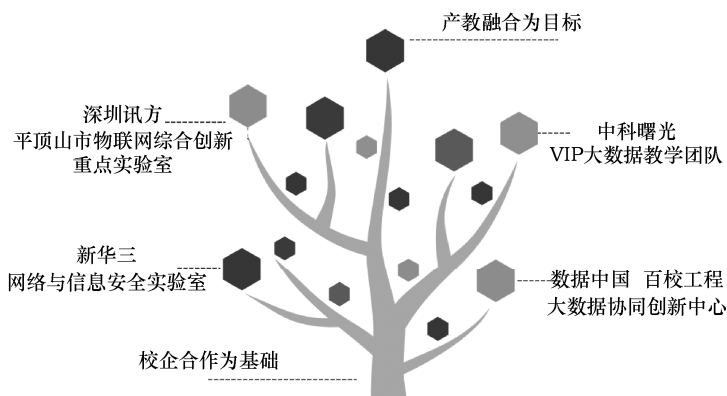


图 2 协同育人平台

（四）协同创新平台建设

按照习近平总书记“万物互联、人机交互、天人一体”的精神和“新工科”的要求，积极推进协同创新平台建设，如图 3 所示。

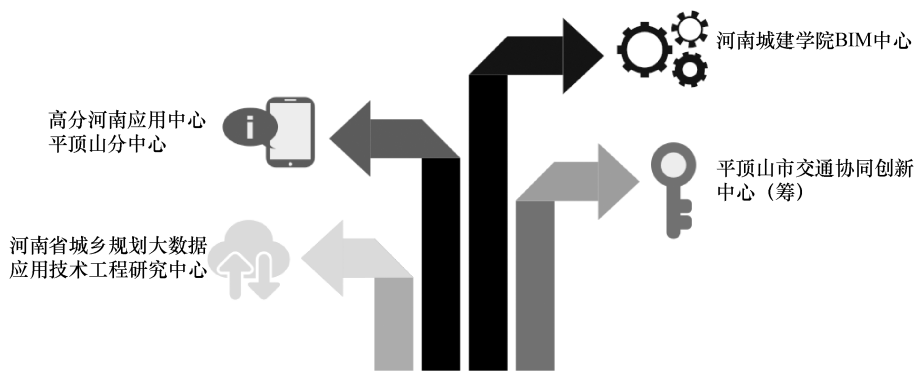


图 3 协同创新平台

（五）大数据协同创新中心建设

投资 1000 万元，依托教育部产教融合创新项目“数据中国 百校工程”大数据应用创新中心。

大数据中心立足于大数据智慧城市建设、面向河南建设领域，服务于智慧城市建设、服务于科研人员学术研究的大数据支撑、服务于大数据人才培养

养，构建协同创新、应用服务和人才培养三位一体的协同创新平台。目前在河南省新型墙体材料大数据、河南省地下空间资源大数据，以及住建行业的数据源建设、应用服务等开展了大量的工作。

三、工作成效

到目前为止，河南城建学院在人才培养、科研创新及应用服务方面做了一些工作。

人才培养：与中科曙光共建的大数据学院已开始启动，首批人才培养方案基本完成，相关的教学工作已经开始，围绕大数据创新创业的双创中心成立。

科研创新：以数据源建设来积极推动数据中心的建设，承建了高分河南中心平顶山分中心，积极推进平顶山市智慧交通协同创新中心的建设，与河南省住建厅签订了战略合作协议，为河南省住建厅提供数据支持。

应用服务：重点启动了河南省墙材大数据项目的实施（该项目已投入运营），与住建部的业务洽谈正在共建中，河南城建学院作为平顶山智慧城市建设的重要参与者，积极参与智能交通、智慧环保等智慧城市的领域建设。

（1）河南城建学院大数据学院建设已经到位，信息管理与信息系统专业（大数据方向）学生培养工作已经开始，基于VIP（Vertically Integrated Projects Program）的师资队伍基本建成，数据科学与大数据技术专业申报工作积极推进。

（2）申报的“河南省城乡规划大数据工程技术研究中心”正式成立，数据源建设、行业应用稳步推进。

（3）依托校企合作、面向智慧城市建设的“平顶山市物联网综合创新实验室”建设到位。

（4）面向城乡规划、房地产、城市地下空间等领域的VR/AR综合创新平台正在积极推进。

四、思考

经过一年多的思考与规划，河南城建学院确立了大数据协同创新中心的发展目标和定位，以大数据学院为基础实施人才培养，将服务与住建行业面

向河南，为河南建设目标的实现提供数据支撑，为河南省大数据改革创新实验做出自己的贡献。以与校企合作为手段，实施协同创新。

在将近两年的运作过程中，也存在一些问题：

（一）人才培养方面

数据科学与大数据技术专业的专业建设规范尚未出台，规范化的人才培养方案尚未确定，系统化的专业教材还需进一步完善。

（二）数据源建设

数据的权威性与真实性之间的差距需要依托大数据技术不断的修正；数据主权和数据使用权矛盾还需通过法律法规的途径不断规范；数据的开放性和数据的可用性之间还有较大的差距，智慧城市建设巨大的数据量的需求和现有数据供给量之间的不匹配还需进一步调整。

（三）协同创新平台还需积极推进

全国大数据教育联盟在大数据人才培养、行业应用等方面做了大量的工作，但大数据技术平台更新周期短、应用技术不断演进、行业应用需求强烈，还需要在大数据应用人才培养的全过程、大数据行业应用的全领域、大数据生态环境构建的全系统开展更多的工作，不断加大协同创新的力度、加快推进大数据强国战略的实现。

（四）大数据浮夸现象的存在

大数据作为时下最热门的一大术语，正在引导政府、企业、行业积极推进技术引领的合作与发展，但对于大数据的落地、大数据真正促进行业快速地发展，以及对大数据应用的实现还有很大的空间。

（五）大数据行业快速发展，技术不断进步

企业的大数据应用引领着行业的进步，因此，在大数据行业中，期待加强大数据的协同创新，特别是对于地方性应用型本科而言，关键在于如何与“双一流”高校进行交流、学习，如何充分利用其人才技术优势带动地方应用型高校的应用创新与人才培养。

五、项目创新点与引领示范点

河南城建学院大数据中心的创新点及大数据学院的引领示范作用主要表现为以下两个方面：

(1) 面向行业。作为全国为数不多的城建类院校，河南城建学院主要立足于环境、交通、规划、地信等行业，与住建部、河南省建设厅、平顶山市各局委联合共建河南建设行业大数据中心。

(2) 立足于地方。不断加强与平顶山市政府的合作，大力推进“智慧鹰城”建设的力度。

在过去一年半的时间内，河南城建学院取得了很大的进步，在数据中心建设、基础应用、人才培养、校企合作等方面均有了一定的收获和提高。但在技术应用、高层次人才的指导及大数据专业课程体系、课程设置、人才培养方案优化等方面，还需要社会各界的指导。

作者简介

何宗耀：硕士研究生、教授，硕士生导师，河南城建学院计算机与数据科学学院院长，河南省青年骨干教师，河南省城乡规划大数据工程研究中心。兴趣方向为智能信息处理及数据科学专业课程教学和行业大数据应用研究。

大数据技术人才培养需要跨越的障碍

合肥工业大学计算机与信息学院教授 胡学钢

一、当前大数据领域人才培养所面临的主要问题表象

随着计算机科学领域的迅速发展和广泛应用，尤其是近几年来，随着云计算、物联网、移动计算、大数据、智慧城市等技术、领域的迅速发展，掌握大数据技术的人才需求数量急剧增长，速度更是惊人。

然而，现实是，用人单位很难找到满意的就业者，高校毕业生很难找到满意的岗位。为此，指责学校培养质量的论调也就多了起来，也引发了许多争论。

如何看待此现象？就人才培养的主体单位高校，以及相关部门来说，如何改变这些局面？

下面就这一方面问题，在简要分析其原因的基础上，给出一个建议。

二、问题分析

高校和用人单位之所以在对人才及人才培养的评价方面产生分歧，有多方面的原因，下面简要列出几点。

（1）客观上，由于高校数量的急剧增长，确实存在部分高校在对人才培养目标的理解与定位方面存在偏差，在办学条件和管理与激励机制等方面存在差距，人才培养质量自然存在问题。

（2）由于计算机科学技术领域的迅速发展，导致许多高校“跟不上”形势的发展，尤其是像对大数据这样的新技术的跟进迟缓。

（3）高校和用人单位对高校人才培养的定位难免有偏差。

一般而言，高校以培养通识型人才为主，要求高校毕业生具备本专业领域的基础理论体系和应用能力，并通过毕业设计等实践性教学活动，在某一方面形成一定的专长。

即使这样，也难保证对就业岗位的任务做到“迅速上手”，更何况像大数据这样快速发展的领域，一般需要有一个学习和适应的过程。

用人单位侧重于选择“能干”之人本身没错，但如果希望毕业生刚到工作岗位就能全面胜任，未免期望过高，因为一个有实力的企业必定在某方面有自己的工作深度。

（4）当前高校大数据人才培养大多缺乏必要的业务“场景”和技术。

虽然应用领域的多样性，使得高校人才培养过程中不可能面面俱到，但通过典型领域、典型示范应用的完整实现，可以系统培养所需要的各项能力。

然而，由于真实特定领域的保密或隐私保护的要求，获得人才培养所需要的真实场景也具有较大的难度。

真实领域典型示范应用知识应当包括业务活动场景、可收集的数据（采集选点和频度、规模、质量）、数据分析的目标及其价值等。

如果缺乏真实场景的训练，容易使学生形成对大数据技术的应用停留在抽象的概念，被动接受任务、技术选择、结果展示等。当有机会面临真实场景时，就难以胜任。

虽然许多高校开展了数据挖掘的研究和应用，但也未必有机会及时开展大数据技术应用，由此而导致对各类新出现的技术的适应训练一定的欠缺。

三、我们的对策

（一）基本出发点

第一，大数据领域需要的人才多样性且规模大，需要更多高校参与才能满足社会需要。

第二，短时间内，大多数高校对大数据人才培养的条件建设都有一定的欠缺。

第三，部分高校、企业拥有人才培养所需要的资源和技术。

第四，整合资源，抱团发展，总比各自单打独斗有优势。

（二）构建基于“共建共享，合作共赢”模式的大数据人才培养平台

鉴于上述基本观点，构建社会化的联合人才培养平台，基于“抱团发展，

共建共享，合作共赢”等理念，探索合作模式，是跨越当前大数据领域人才培养壁垒的重要途径。

其核心思想如下：将有合作意愿、有支持条件的高校、企业相关部门建成一个专题联盟，共同探讨人才培养的相关研究、建设和推进，共同建设和分享教育教学资源，从而形成整体发展的局面。

（三）安徽省高校大数据教育联盟的建设

安徽省高等学校计算机教育研究会，是专门从事高校计算机教育教学研究的学术社团，已经成功地开展了许多专题研究、大学生竞赛等活动，并组织了约十个专题联盟或专题组，为全省高校计算机教育教学事业做出了积极的贡献。

“安徽省高校大数据教育联盟”是研究会设立的一个以“大数据人才培养和资源建设”为主题的专题联盟，联合省内高校、相关企业开展以下几方面的工作：人才培养模式研究；教学资源 and 平台建设；新技术培训；大学生大数据应用技术竞赛；教材建设等。

2017年4月21日和22日，与全国大数据教育联盟联合举办了大数据人才培养研讨会，其间，讨论和确定了联盟章程、大学生大数据竞赛方案等。

经过几个月的工作，已经取得了初步的成果。

（1）联盟建设：全省已有27所高校参加了联盟，并有国内几家从事大数据技术的企业加盟。

（2）竞赛推动：研究了“大数据技术与应用”竞赛方案；2017年8月举行了“大数据技术与应用”培训，受训师生153名；11月11日成功举办了竞赛，参赛队数81支，来自全省27所高校；建设的竞赛技术委员会将成为竞赛的核心力量。

（3）人才培养建设与研讨机制建设：目前省内多个高校在申报相关专业或者设置专业方向，并已经开展了多个不同规模和侧重点的研讨，以及相关的建设。例如，合肥工业大学设立了“大数据+创新创业基地”，提供3200平方米的面积，一期投入1500万元，将在建成后面向全校师生开展相关的教学、科研和人才培养工作。

四、结束语

以上这些观点和方案，是基于安徽省高等学校计算机教育研究会近十年的研究和实践的总结基础上所做出的探索，还相当不成熟，希望得到各方面的指导。

联盟的建设还存在诸多影响因素，包括理念上的偏差、经费的使用、成果的共享、投入产出比、平台的管理和可持续发展等。

所有这些，需要积极开展研究，并借助某些平台或资源，方可形成可持续发展的局面。例如，借助政府部门的政策引导或规定，借助成熟的社团或平台等。

作者简介

胡学钢：合肥工业大学计算机与信息学院教授，“数据挖掘与智能计算”千人团队博士生导师。曾任合肥工业大学计算机与信息学院副院长，合肥工业大学宣城校区管委会副主任等职。现为安徽省高等学校计算机教育研究会理事长，教育部计算机类专业教学指导委员会委员，全国高校计算机教育研究会常务理事，CCF 教育专委会常委。研究方向：数据挖掘、知识工程、算法分析与设计。

大数据研究要注意的两个问题

中国人民大学信息学院教授 陈 禹

大数据的研究和应用已经引起了广泛的关注，这是值得高兴的。但是如何科学地、有效地开展大数据的研究与应用，是值得我们思考的。冯·诺依曼说，现在的经济科学缺乏大数据，缺乏海量的数据分析。现在的数据还不足以研究我们面临的所有问题，我们要将注意力放在数据的收集上，但在同一篇文章中，冯·诺依曼明确指出了，对于概念尚不明确的数据进行大量的计算是毫无意义的。

和冯·诺依曼的时代相比，我们今天的数据要多得多，但冯·诺依曼指出的理论思考与数据分析的有机结合仍然是我们值得注意的问题。有一个说法认为，只要有足够的数据，具体领域的知识、具体的思想方法、具体的理念都已经不重要了，甚至有的说法认为，逻辑关系和因果关系都已经不重要了，个人认为这种观点是不正确的。实际上，近年来已经看到了很多事实，如“黑天鹅事件”很多，很多统计分析得到的结果与现实差别很大。因此，经常用到马太效应、肥尾效应等词，不论在自然科学中，还是在社会科学中，现有的统计方法具有一些根本性的弱点，并不是说统计方法不重要，而是要知道统计方法隐含的前提便是所谓的同质，我们事先要假设所有的统计数据是面对同样的实体。但现实比任何理论都要丰富得多，我们很难要求现实中的多样化实体遵从完全同样的事先假定的同质化。正因为这样，很多分析往往和现实相去甚远，这种情况在近几十年越来越多。进一步来讲，可以看到，问题出在研究的基本理念上。

一、大数据更需要科学的思想方法

今天包括理工科和文科，几乎所有学科的思想方法都是近代科学思想方

法。其主要弊病在于将客观世界过分简单化、过分同质化，没有看到客观世界的复杂性，因此，当其理论用到现实中时，往往会和现实出现种种差距。这种情况在 100 多年前由爱因斯坦、普朗克等伟大的科学家首先发现，这 100 多年来，科学家们在各个领域，越来越多地发现原有理论的同质化、简单化所造成的诸多问题。这种情况越来越多地引起了学者们的注意，因此，目前在学术界已经出现了一个新的研究潮流，即所谓复杂性研究的研究趋势。

二、复杂性研究

复杂性研究并不是一个具体的学科，而是一种思想方法，这种方法强调的是承认世界的复杂性，重视客观事物的质的多样性、质的无限性，正是我们面对的种种不确定性。这方面著名的学者有赫伯特·西蒙和约翰·霍兰，他们在 20 世纪末对于复杂性研究的基本理念进行了系统的叙述，得到了各学科的广泛注意。因此，我们在研究大数据时，首先要摆脱近代科学所造成的束缚，要认识到客观世界不仅在量的意义上是无限的，而且在质的意义上也是无限的，特别是层次的概念，当跨越一个层次时，其规律、现象都会产生新的变化，这种现象在学术上称为涌现。

约翰·霍兰有一本书叫《涌现》，对于各学科中出现的涌现现象进行了深入的比较，而在赫伯特·西蒙的《人工科学》一书中，进一步将客观复杂系统的层次性进行了深入的分析。其实，复杂性研究的出现很大程度上便是我们经常讲的系统科学、系统工程方法的进一步深化，大家都知道一加一大于二，对于为什么大于二、怎么大于二还远远未搞清楚。因此，当我们在用一些传统的统计方法加工数据时，都是按照一加一等于二的思路，而客观世界是一加一大于二，差距便由此产生了。2017 年的两个突出事件便证明了这一点，一个是美国大选，另一个是英国脱欧。并不能说这些调查分析机构的分析少，也不能说其计算方法不正确，但事与愿违，出现了所谓的“黑天鹅事件”。除了其他的种种原因之外，很重要的一点是他们对于现有的统计方法、计算方法的局限性没有客观的认识。

因此，大数据研究中的两个问题是一件事的两个方面，一方面，我们要对于所有的分析方法和分析理论框架有一个不断改进、不断扩充的思维方法，

绝不能认为一种理论能放之四海而皆准，在任何情况下都能适用。比如牛顿力学，并不能说牛顿力学不对，也不能说牛顿力学在任何情况、任何尺度、任何问题中都是适用的，爱因斯坦和普朗克的贡献恰恰就在这里。另一方面，我们现有的统计方法都拿来做数据分析、数据研究，这无疑是必要的、重要的，但我们要明确目前所说的数据方法的局限性和前提，因此，和冯·诺依曼的时代相比，我们今天已经有很多数据，已经能有很多方法去收集以前收集不到的数据。这是和前人相比比较幸运的地方，但大数据绝不等于不要理论、不要思考。

还有一个相关的问题是我们对于各个领域的特殊性认识，我们常说隔行如隔山，每一个领域有特殊的知识和规律，但我们将大数据方法用到某一个领域时，一定要将该领域的特殊性放到首位，要深入地了解 and 掌握实践所提供的丰富材料，认识到在这个领域中应用到大数据的分析方法应该要注意的与其他领域不同的特点。这也可以说是强调实践的观点，我们说，研究大数据要顶天立地，顶天是指对世界的了解要有现代的信息技术、掌握尽可能多的数据，立地是指紧密结合具体领域的问题，如经济问题、金融问题、电子商务问题等。

大数据的研究和应用是非常具体、非常实际的，很难笼统地说应该怎么做、应该给学生教些什么，但我们如果对于理念与方法有一个科学的、比较统一的认识，便可以对这个领域的拓展做出应有的贡献。

三、方法问题和教学问题

大数据技术目前有很多方法，但用得最多的还是统计方法、概率论等，这些方法各有千秋，各有各的用处。作为大数据的研究者，对于每种方法都要进行认真的研究和分析，包括分析结果的可视化，这是很具体的技术问题，作为大数据的研究者，要将所有的东西作为我们的工具箱。工具是重要的，而且是越多越好，各有各的用处，但不要以为一种工具能解决所有的问题。与此相关，便是教学的问题，我们要培养的大数据人才应该是能在各个具体问题中发挥大数据的作用。因此，对他们而言，一方面要理解世界的复杂性，准备应对各种各样不同的质的问题；另一方面要掌握尽可能多的工具，用适

当的工具解决适当的问题，这便是大数据研究中需要注意的两个问题。

作者简介

陈禹：工学硕士，教授，博士生导师，经济科学实验室主任，曾任信息学院院长，信息系主任。编写的教材与专著有《复杂性研究视角中的经济系统》《CAS 理论及其应用》《信息经济学教程》等。国外有关领域的经典著作翻译：在进行的有《社会网络分析》《实验经济学经典论文集》等 5 本。近一年内有国际、国内学术交流会议 3 个。主要研究方向有计算机应用、信息管理、系统科学等。参加的研究项目包括设计模型、编制软件、研究理论，在此基础上撰写论文。主持和参与的科研项目包括资源运营的模式与支持技术研究（国务院发展研究中心项目）、国家电子商务白皮书的编写（商务部项目）、教学数字化环境研究（教育部项目）。曾任社会兼职中国信息经济学会理事长，国际信息系统学会中国分会副理事长。

对大数据专业人才培养的几点思考

上海工程技术大学电子电气工程学院副教授 黄润才

2017年，上海工程技术大学作为首批招收数据科学及大数据技术专业的高校之一，第一批学生已经入学。这是国家或教育部对于学校前期工作的认可，也是对后续工作的鞭策。作为参与申报组织相关工作的老师，谈一下近些年的体会及目前开展教学的体验。最要紧的是，不能将大数据专业定位成计算机学科的大数据方向。由于大数据的体系有自身的技术线路，我们在招收工作进行之前做了大量的调研和研究。

一、理解大数据专业

（一）专业特色

大数据专业是一门交叉性学科，以统计学、数学、计算机作为三大支撑性学科，以及一些应用领域作为拓展性学科。因此，大数据专业的培养体系比较复杂，且口径比较宽，这时是以宽口径、重应用为主的特色，包括数学基础、知识体系、编程能力、平台运维、系统架构和应用开发能力、算法研究能力等。因为大数据目前由政府与行业主导，高校还未直接掌握大数据资源。因此，在大数据办学的过程中，应该是政企产学研合作结合的方式，政府支撑、高校主导、行业资源利用、科学院所联合开发平台，共享互动，联合办学。

（二）应用领域

大数据应用形势发展非常迅猛，在此背景下，本科大数据专业近几年纷纷被高校申报。目前，大数据应用在各行各业取得了突飞猛进的发展，但教育行业还未把握大数据的形势，然而教育却站在培养大数据人才的最前沿。金融、政府部门、电信、零售、安全、交通、媒体等行业处于大数据应用的

前沿，也是这些行业在推动大数据人才的应用。

各行各业在大数据的获取和应用领域中，教育在目前状态下在应用数据中不具有优势，但具有教学资源、教学平台、开放式办学理念，在人才培养方面要走出自己的一条独特道路，与政府企业相融合。

大数据的应用模式主要有 4 个方面：数据源、数据获取与存储、数据分析与应用、可视化产品与服务，这 4 个方面也是培养人才过程中的重点教学内容，其中每个过程都有可能涉及智能化体系，人工智能的结合是大数据的必然方向。大数据应用可以总结为两大类：精准化定制服务和决策支持与预测。所谓精准化定制服务是指供需双方依据大数据找到对需求方的特定化服务，通过大数据的按需分析达到最佳配置，类似于定向购物。通过大数据的分析提供定制化服务，达到精准化的方向。决策支持与预测是根据过去的数据或大数据容量，加入相关的因素，就某个对象进行数据分析，能提供预警或动态优化，做出一些优化的结果对决策支持提供方法，包括风险预警和实时优化，

（三）人才的需求

各行各业都需求大数据领域的人才，从目前的结业形势来看，很多企业无论大小都在需要大数据行业领域的人才。

二、大数据相关技术

不能把大数据专业定位成计算机学科的方向，作为一个专业，肯定有知识体系、基础、应用、平台。

图 1 所示为大数据相关技术。



图 1 大数据相关技术

（一）数据获取

大数据如果作为一个行业，数据获取是处在前沿的一项技术，通过各种数据获取的手段，能获得大数据的数据来源。物联网是前端获取数据，包括社交网络、移动互联网等，这些是获取数据的主要来源，尤其目前的用户越来越多样化，数据的品种和形式也越来越多样化，结构化、半结构化、非结构化的海量数据都是大数据的来源，由于成千上万的用户可能同时进行访问和操作，因此，要对大数据采集方法进行一些分类处理。

网络的数据采集是通过网络爬虫、网络公开 API 方式从网站上获取数据信息的过程，将结构化、半结构化及非结构化的数据从网页中抓取出来。流量的采集可以采用 DPI、DFI 等带宽获取技术进行处理，这也是大数据的传统来源，目前也是依然存在的。因此，大型数据库（包括网络后台的大量数据）是用来存储数据的，通过这种方式采集端部署了大量的数据库，如何在数据库之间负载均衡和分辨进行深入的思考和设计，这些都是数据库存储大数据需要考虑的问题。

（二）数据存储

在进行数据存储时，传统的计算机学科会提到数据库，但大数据存储数据可能不仅仅是单纯的数据库问题。这里指的存储包括用户端（数据来源），应用服务层是计算机学科中经常遇到的概念，在传统的计算机学科中讲的是数据库，但到了大数据存储时，数据存储服务层可能会有大量的体系结构。

（三）数据可视化

在后台如何呈现给用户或如何通过数据分析将用户需要的结果以图形化、图像化的形式表现出来，这是大数据中很重要的一个方法。可视化主要是借助于图形化的手段清晰、有效地传达与沟通信息，将大型数据集中的数据以图形图像形式表现，并利用数据分析和开发工具发现其中未知信息的处理过程。在进行大数据可视化分析设计时，有以下几点需要在未来的教学工作中多加注意。

（1）可视化分析：可视化的显示结果根据用户的需要可以定制一些显示形式，如以地图、树状图、柱状图、区域图、雷达图等方式，将后台大量的

数据分析结果展现在用户面前。由专业人员开发一些显示工具，或利用现有的图像化、可视化工具将数据接口做好。

(2) 可视化信息：关注知识的可视化展现和图形的设计。

(3) 可视化工具：如何将数据进行可视化，这一点以前在计算机学科中有很多可以使用的工具，如 Excel、MATLAB 等都具有可视化的功能。传统的工具可能处理的数据量比较小，展示形式也不是很丰富。在大数据领域中，最重要的、技术手段最高的、发展最为迅猛的也是数据分析工具，也是大数据人才培养方面技术含量最高的一个方向。

(四) 数据分析

提到数据分析，传统的计算机学科会与数据挖掘进行对比，也是数据分析的一个主要方法，但与大数据分析相比有各自的特点，比如算法的复杂度、数据的状态、在线的或存量和增量的关系。大数据、云计算、云存储对数据和环境的要求不是很高，且大多是结构化的数据。

在数据分析方面，包括实时数据分析、可视化数据分析、智能化数据分析、文本数据分析、Web 信息分析等，这些可能是当前比较流行的数据分析方式。

三、人才培养的思考

高校老师考虑最多的可能是如何形成一个培养体系，如何将学生培养好，无论是教育人员还是科研人员，都需要在系统的培训下形成一个完善的知识结构。从企业、应用掀起的一个浪潮，高校的培养往往会滞后于企业的需求，这给高校带来了契机，同时也带来了很大的压力。

(一) 培养层次

大数据的人才不一定全部是高学历，技术从底层到高层算法的研究，其实是有递接性质的，因此，大数据人才培养也一样，从专科、本科、硕士、博士可能都会找到自己的就业岗位。在培养层次方面，可以是多方位的，目前高校中一般是本科、硕士或博士的培养层次，如图 2 所示。



图2 大数据人才的培养层次

（二）培养体系

教学计划或培养方案应该是从底层到高层形成一个金字塔的结构。从底层开始依次为新工科建设、信息学科大平台，包含数理技术、计算机技术、大数据的专业知识，以及本科生和研究生的培养。研究生阶段可能更侧重于一些高阶阶段的算法研究。

（三）总体规划

上海工程技术大学对大数据人才培养的方向上做了一些思考与规划。全校已有的学科资源结合优势学科、优势力量集中办好大数据，当然是以电气工程学院为主，计算机学科主办的模式，以及数理统计学提供数学基础方面的支撑。一些比较强的学科已经开始运用行业特色做一些行业应用，如城市轨道交通、航空运输、服装设计等。集中优势学科能拓展学科平台，办好大数据专业，将优势学科集中起来，以计算机学科为主，考虑各个学科的优势和应用，主要是一些平台的应用。

四、进一步的工作

（一）校企合作

校企合作是大数据专业的必经之路，目前很多学科都是产学研相结合，大数据专业尤为突出。近几年，上海工程技术大学先后与很多行业与企业开展了产学研合作交流，包括国外的华盛顿大学、南洋理工大学、苹果公司、甲骨文、上海仪电集团、上海未来宽带技术等国内外高校和企业都纳入这个平台中。

（二）教师队伍

在申报大数据专业时，已规划了师资队伍的建设，具体包括教授 7 人，副教授 10 人左右，讲师有一批形成梯队，其中包括刚毕业的博士及拥有博士学位的 20 人左右，整个队伍在构建大数据专业的培养体系时已形成了相对完善的师资队伍，但相关的队伍仍有待加强，如挂职、在职培训。结合大数据行业的发展形势必须走出校门，与企业、政府挂钩。

上海工程技术大学在人力、物力、财力方面对大数据技术及专业的筹建给予了大力的支持，大数据专业逐渐形成了机器学习与计算机视觉、智能计算与现代交通信息融合、绿色计算与大数据应用、信息物理融合与物联网工程这四个具有鲜明学科特色和发展潜力的学科方向，为数据科学与大数据技术专业的开办和进一步发展提供了学科支撑。

（三）实验室建设

本校与中国最大的大数据操作系统厂商联合成立了大数据实验室，以工信部全国云计算及大数据应用技术人才培训考试认证为契机，合作开展大数据分析人才的培养工作。先后建设了与华盛顿大学共建的联合实验室，与企业共建的视觉研究所等，这些研究中心为大数据教学的开展奠定了基础，同时也提供了很多实习的机会。

作者简介

黄润才：上海工程技术大学副教授，硕士毕业于华中科技大学大学计算机学院，博士毕业于东华大学信息学院，华中科技大学上海校友会人工智能分会会长。主要研究领域：智能计算、计算机网络及应用。



产业篇

新时代下的互联网产业变革



“互联网+”背景下的企业转型与变革

北京大学光华管理学院副教授 董小英

一、转型和变革的界定

转型和变革是两个不同的概念。转型通常是指在宏观环境中、技术环境和市场环境中出现了新的要素，行业发展态势和商业模式发生改变，企业要针对外部环境的变化对其发展战略进行重新调整和定位，以上界定为转型。在外部环境发生变化的同时，企业内部也需做出调整和改变，包括原有的思维模式、组织体系、业务模式及业务流程，因此，转型主要是针对外部环境变化进行战略调整和修正；外部变化驱动企业内部资源配置和人、财、物战略布局调整称为变革。

二、“互联网+”转型的挑战

在互联网时代，年轻企业以极快的速度在颠覆一些传统企业的做法。他们通过快速的学习能力、迭代能力、勇于试错的能力及承担风险的能力，不断在新兴领域快速形成大型的互联网企业。由年轻人主导的企业所带来的改变主要发生在边缘、脆弱行业和松管制行业。这类行业中，国有企业相对较少，产业集中度较低，缺乏极具竞争力的领先企业，即使有领先企业，这些企业也不会将技术作为驱动企业发展的关键要素，如京东、阿里巴巴及滴滴均是如此。在这些行业，准入门槛相对较低、获客机会相对较大，新兴的互联网企业通过快速迭代、客户体验优先、相关行业交叉渗透和逐步扩散等模式寻求快速崛起路径，给传统企业带来了很大的转型和变革压力。

三、竞争空间的改变

在信息技术领域，人们通常将其翻译为信息物理系统，但我们将其称为

网络—实体空间系统。在实体空间（P），中国对资源竞争已经非常激烈，很多传统行业，如房地产行业、零售行业等在实体空间的竞争已趋向饱和状态。在网络空间（C），凭借互联网技术进行数字化和智能化开发才刚刚开始，预判到这个空间价值的企业将获得未来的竞争优势。中国和美国的互联网企业从整个社会格局及全球视野来看都在此进行战略布局，发展在网络空间整合资源的能力。在网络空间的能力构建并不是放弃实体空间的资源，而是以更宽阔的视野、更长远的战略思考和更强的知识及技术能力实现对实体空间资源的整合和配置，减少由于供给侧与需求侧信息不对称、信息延迟和信息缺失所带来的错配（如滴滴便是一个典型的例子），在社会层面整体优化资源的利用效率。因此，系统能力（S）是在实体框架和网络空间建立联系，增强互动的能力，通过整合信息流来合理设计利用业务流、资金流和物流，商业的本质并没有改变，但是，掌控商业的路径发生了变化，传统企业擅长在实体空间整合资源，而互联网企业从网络空间整合资源，未来的趋势仍然是网络与实体空间的融合，但关键是看谁的速度快、能力强。

四、主导逻辑的改变

在这个转型与融合的过程中，主导逻辑决定了人们的战略布局与资源配置。所谓主导逻辑，是指人们做生意的一种思维模式、认知方式。传统工业中企业管理者是以自我为中心的，眼光聚焦在自身能力、资源上，主要考虑我能够生产什么产品，产品体系做好后，将产品推向消费者。也就是说，先产品，后市场推广。对传统制造业而言，主导逻辑是以生产者和产品为核心的，由于工具和手段的缺乏，这些企业要想了解客户需求、获得客户反馈，成本和代价是很大的。因为很多消费者对企业产品的使用无法留下痕迹，企业也无法得到反馈。在市场稳定、产品短缺的情况下，产品主导逻辑是非常有效的思维模式和业务流程，持续的产品改善和流程优化带来了竞争优势。但是，当市场剧烈波动或竞品大量涌现时，供给侧的生产方难以准确预测需求方的变化，在应对外部市场动荡时只能盲人摸象，自乱阵脚。

但在今天，大多数互联网企业通过搭建平台聚焦了海量客户资源，互联网企业以高效手段及低成本获取和分析消费者信息，为解决供给侧与需求侧

的错配问题带来了可能。服务主导逻辑大行其道。服务主导逻辑最核心理念是以消费者的需求为中心，抓住了需求就抓住了供给，产品不再是主角，而是为满足消费者需求、提供服务的一部分。通过互联网平台模式，供给侧与需求侧双方成为彼此之间快速修正、快速调整和快速匹配的双轮，厂商与消费者的距离从来没有这么近过，交流从没这么快过，因此，虽然商业的本质并没有改变，但是商业运作的效率大大提高了。在“互联网+”和服务主导逻辑时代，传统企业在转型过程中需要明确下面几个变化：

（1）消费者与企业共同创造价值。以前的企业是闭门造车，如今越来越多的企业采用开放式创新，不论是众筹，还是消费者需求调研方式，越来越多的消费者介入到产品研发的过程之中。

（2）产品是服务的一部分。服务是针对消费者需求的一种全面的解决方案，越是能完整地满足消费者需求的企业，其在产业龙头中的地位越重要，这是一种互联网环境下产品主导逻辑向服务主导逻辑转变的一大标志。因此，在这个市场环境中，谁能抓住消费者的需求，谁能全面满足消费者的需求，谁便是整个行业的龙头。

（3）整合资源者是关键中枢。就一些互联网企业而言，如滴滴，大部分出租车并非其拥有，而只是调配、利用及客户反馈、使用出租车的资源，在网络空间具有整合资源的能力，这类企业在如今也是最具竞争力的企业。

（4）操纵型资源是创造价值的战略资源。如果我们在网络空间可以整合数据资源，通过大数据分析、机器学习、深度学习掌握的算法，对这些资源进行深度开发，了解消费者的行为偏好，根据其需求进行相应的市场营销活动。因此，掌握操纵型资源的企业成为“互联网+”环境下的领导企业。操纵型资源一方面可以整合实体资源，另一方面也是一种分享经济模式。未来这种模式会逐渐从租房行业、出租车行业向其他领域扩散。但发展分享经济的关键能力是在网络空间具有整合能力，整合的数据具有开发这些数据的操纵型资源，其中不仅包括实体资源，还包括一些较弱的操纵型资源。

五、传统企业的转型

总而言之，对绝大多数传统制造型企业而言，“互联网+”并未改变商业

的本质特征，但延展了传统行业中原有制造业和服务行业中客户交互比较弱的环节。互联网企业将客户信息挖掘转化成企业研发、产品和服务创意的来源，它们成为企业可持续发展的战略性资产，这种与客户的深度交互和连接，使得企业和客户不分彼此，成为协同创新和共生的沃土和有机体，这样的企业是不会被市场和消费者抛弃的。

如果传统企业有较好的信息化基础，可以在消费者的交互端强化以前缺失的能力，将与消费者交互的数据变为战略资产，通过消费者的数据分析，指导和调整生产计划和运营体系，使其更加满足运用市场和消费者的需求，变为真正的柔性化和高适应性企业。

但对于已有的内在信息化能力构建速度和完整性相对薄弱的企业而言，仍需要坚守定力。一方面，传统制造业应坚信好品质的产品一定会有消费者和市场。另一方面，企业要在自身的信息化构建上做更大的投入、更快的成长。将信息化作为自己战略优选项，构建自己的客户群，搭建与消费者之间的数字化协同、数字化交流、数字化共创价值的能力，只有这样才能在“互联网+”大环境下获得新生。

作者简介

董小英：北京大学光华管理学院管理科学与管理信息系统系博士生导师，副教授，案例研究中心学术主任，中国信息经济协会副理事长，工业和信息化部通信专家委员会委员，国家知识管理标准委员会委员。曾在美国哈佛大学、匹兹堡大学、澳大利亚国立大学等多所大学短期学习或做访问学者。主持联合国教科文组织、国家自然科学基金和国家社科基金项目。主要著作 6 部。发表中英文论文 50 余篇。研究教学领域为企业知识管理与创新、企业信息化战略、德国工业 4.0、数字化时代的组织转型与变革、竞争情报等。对中关村企业、思科、华为、京东、腾讯、李宁等企业有深度案例研究，担任多家企业信息化与知识管理顾问。授课内容包括知识管理与创新、变革管理与商业模式创新、领导者创新思维、企业信息战略等课程，对象包括 EMBA、EDP 和 MBA 学员，擅长设计和指导针对企业转型变革的行动学习。

大数据与企业管理决策

北京邮电大学软件学院副教授 傅湘玲

关于大数据在企业管理中的应用我们要回答 4 个问题：第一，这些数据说明了什么问题；第二，这些数据从哪里来；第三，我们得出了什么分析结果；第四，在结果中得到了什么启示。要实现大数据在企业管理决策中的应用，一方面要有好的数据支撑，另一方面需要经典的管理理论的应用。只有数据与经典理论结合起来，才可能形成新的管理决策的应用和模型，这是笔者对大数据与企业管理决策的理解。

这里分享 5 个案例：第一是基于海量的互联网数据的新产品开发决策；第二是基于海量互联网数据的竞争产品分析；第三是基于企业社交网络的员工潜力测量研究；第四是利用公众博客文本进行公众幸福感测量；第五是基于微博数据的新闻线索发现。

一、基于海量的互联网数据的新产品开发决策

传统的新产品设计一般是通过问卷的方式进行的，用户买了产品之后，会留下很多评论，这种评论实际上代表了用户的需求，我们能否将这些用户需求转变为产品设计元素，从而改进产品设计？我们能否将经典的卡洛模型用在在线评论分析之中，从而实现智能、及时地实现新产品的改进？如针对手机产品的评论所做的二次开发过程中，手机新产品开发过程中如何利用在线评论提取其需求，从而帮助设计师更好地改进产品设计。图 1 所示为技术路线图。

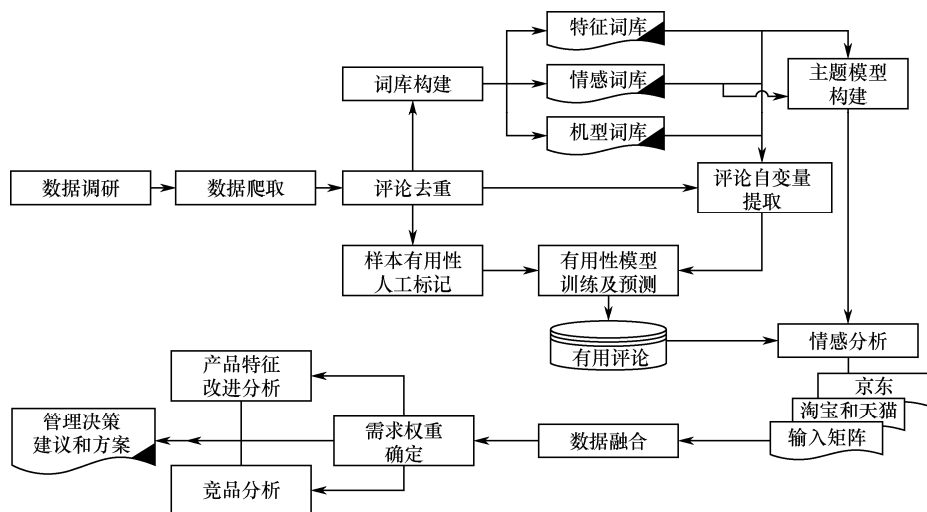


图1 技术路线图

在上述技术路线图中，首先做一个数据调研，从京东、淘宝、新浪微博中提取我们需要做的手机评论数据，本研究选取了10款需要分析的手机型号，从京东等网站上进行数据的爬取。获取到数据之后，在技术路线图中可以看到进行了数据的预处理，其中包括评论的去重，当然还有一项很重要的工作，即样本的有用性人工标记，其实对产品设计师而言，有些评论对产品设计师没用，但对消费者有用。做完样本的有用性标记之后，进行有用性模型训练，同时在大量的评论中构建一个需要提取的特征、情感、机型，因此，在技术路线中，构建了特征词库、情感词库、机型词库，在此基础上构建手机的主题模型，主题模型是指构建一个词对，如手机的待机时间较长。接着进行情感的分析。做完这项工作后，再结合管理中的卡洛模型进行客户需求分析。卡洛模型中提到客户的满意度与基本需求、期望需求和惊喜需求相关，我们根据用户效用值的大小进行排序，得到用户的①基本需求：版本、功能、外观、物流及售后、其他；②期望需求：处理器及配件、屏幕、信号及发热、相机；③惊喜需求：电池、价格、手感、系统。对此我们也提出相应的管理建议——对于基本需求的管理建议：保证符合服务标准，努力降低产品故障率和服务失误率；对于期望需求的管理建议：不单是考虑符合服务标准，而是如何提高服务标准；对于惊喜需求的管理建议：首先保证另外两类需求，开发新服务，增加新内容。

二、基于海量互联网数据的竞争产品分析

在产品评论中存在不同产品间的各种不同属性特征的比较，在此基础上我们提出了另外一个概念——产品在线声誉。产品的在线声誉分为产品美誉度和知名度。美誉度又从属性美誉度和属性权重两个角度进行考虑。就属性的美誉度而言，前面的过程中提取了手机的每个属性特征，如电池、屏幕、内存等，对每个属性都有一个评价的矩阵值，即一条评论中对某个属性的效用值，据此计算出属性的美誉度，接着对属性的权重进行计算，便可得出第*i*条评论对某个产品属性*j*的评价，从而测量出不同产品的在线声誉。在此案例中，我们针对4款手机进行了研究，分别为华为、iPhone、三星、联想。

得出的结果如图2所示。

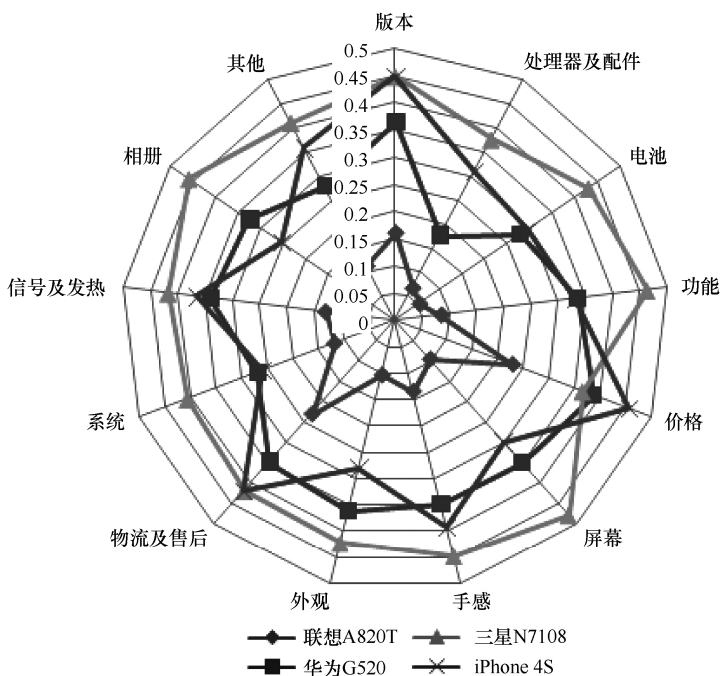


图2 手机产品美誉度对比

- ①三星 N7108 的各个属性的美誉度均在最外围（价格除外），即美誉度高；
- ②联想 A820T 的各个属性的美誉度均在最内层，即美誉度最低；
- ③iPhone 4S

的大部分属性的美誉低于三星 N7108（价格除外），却高于华为 G520（相机、外观和屏幕除外），因此，从属性美誉度层面来看，三星 N7108 表现较好，最好能再适当下调点价格，联想 A820T 整体上都需要提升。

三、基于企业社交网络的员工潜力测量研究

人力资源管理中企业员工的潜力研究一般基于问卷进行，在这里我们希望通过企业内部社交网络的数据来进行员工潜力的测量研究。我们选取了某企业社交网络中员工的社交数据，在此基础上将员工的潜力分为协调潜力和知识潜力两个维度。在此基础上进一步构建每个细化的指标测量方法。通过对文本数据的分析与挖掘，量化测量出每个指标的值，从而进行员工潜力指数的测量研究。

四、利用公众博客文本进行了公众幸福感测量

能不能利用文本进行幸福感的量化测量呢？传统的做法是 Watson 教授提出来的 PANAS 量表，通过问卷的方式测量某个人的幸福感。但这种量表的方式无法实现大规模、可重复、无干扰的测量，也就是说，很多人在测试时未反映出真实的感情。因此，要实现无干扰环境下大规模、可重复的测量，则需要一个更好的可以利用海量客观数据的自动化方法测量公众的幸福，我们做了一个测量幸福感的模型，主要是分析从某一篇博文中出现的情感词数量及频率在整篇文章中所占的比例。其中有一个很大的问题，即中文的情感词库需要量化，传统的词库很多只有正面和负面，对每一个情感词并没有得分的比较，这是工作过程中很大的一个难题，英文中有公开的词库，经过多方努力，我们找到了 Ren 词库。

可以看出，模型的结果与实际情况是比较符合的，我们对历史已经发生的事件和现在模型的结果对比是可以对应的，图 3 是我们对公众幸福感利用博客文本做的结果和重大事件的对比。同样，我们也做了周、年的比较（图 4），将 6 年中每年的数据进行对比后发现，每年的 2 月是情感较高的，由于 2 月有春节，春节后幸福感开始下降，同时“十一”也是如此。在周的对比中，周一较低，周二较高，由于工作比较疲惫，周三比较低，由于看到周末了，

周四之后又开始上升。这是关于重大事件的对比、每年高峰及低峰的对比及一周的对比。

因此，在这个研究中，我们将经典心理学的主观幸福感测量（PANAS 量表），利用互联网中大量非结构化数据设计了一个新的幸福感量化模型，实现了对社会公众幸福感的实时动态监测。

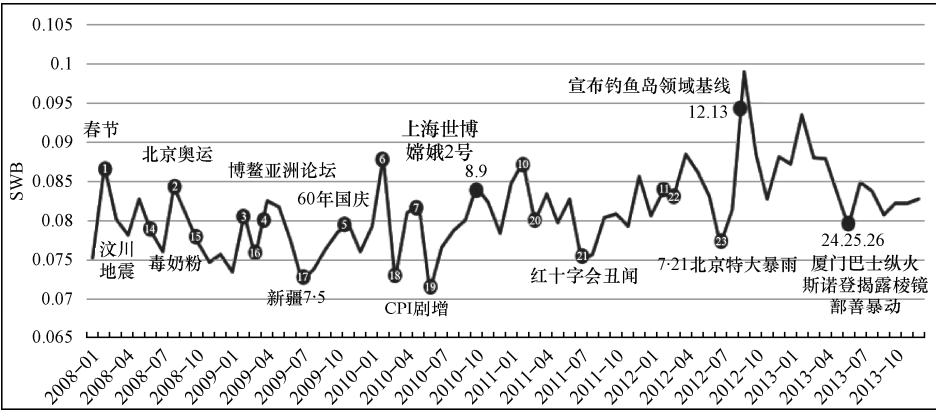


图3 2008—2013年中国公众幸福感变化与重大社会事件对照图

图4所示为2008—2013年中国公众幸福感变化对比图。

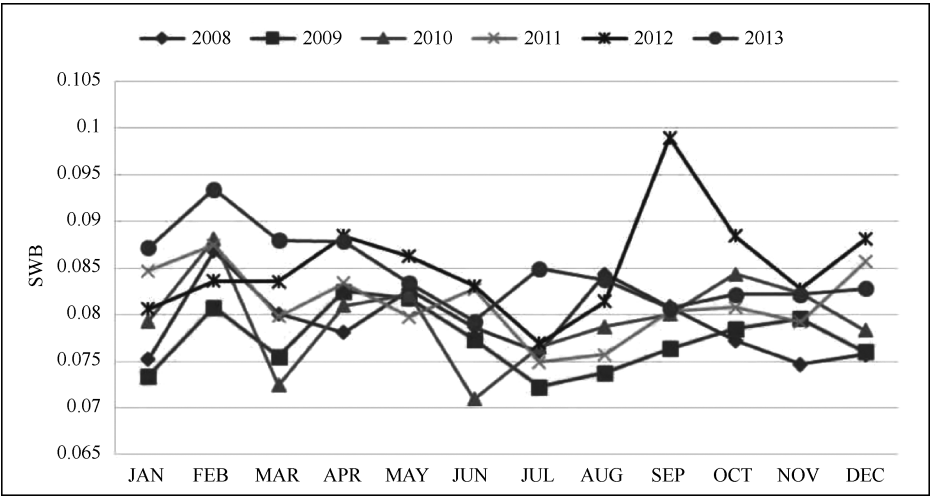


图4 2008—2013年中国公众幸福感变化对比图

图 5 所示为中国公众幸福感一周变化趋势图。

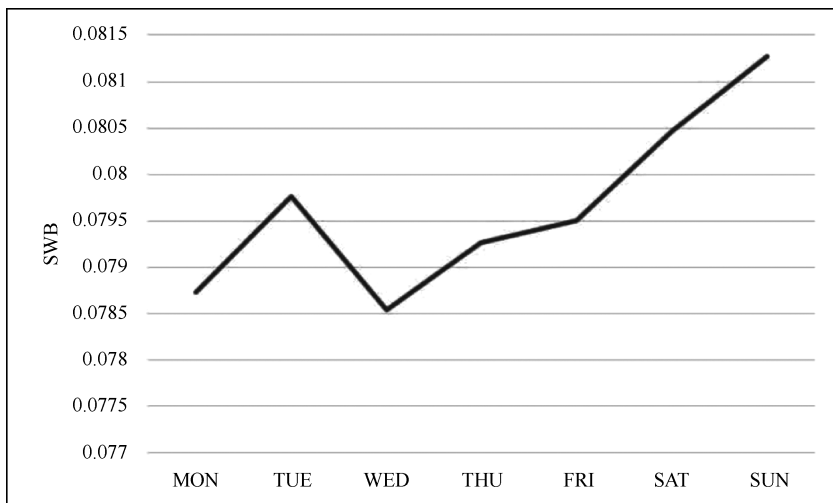


图 5 中国公众幸福感一周变化趋势

五、基于微博在线数据的新闻线索挖掘

目前来看，记者也是通过博客、社交网络发现大量新闻线索，如通过微信群、QQ 群、微博等发现有哪些热点，根据自己的知识判断，这有可能是一个值得深究的会成为一条新闻的消息，在此过程中可能浏览过一万条微博才发现一条值得调研和采访形成新闻的内容，我们称为新闻线索。首先，我们构建了一个新闻线索的新闻价值模型，提高了线索的重要性、异常性和权变性。在构建了新闻线索后，我们听取了新华社、人民日报的记者，以及一些新闻专家、公众的看法，进行了模型的改进，在技术路线图（图 6）中可以看到，一方面是构建新闻价值线索模型，另一方面是从数据中找到新闻线索，在数据准备阶段，主要利用微博对事件进行了事件触发抽取、命名实体识别、时间表达抽取、事件后果抽取，由于在新闻价值模型中发现，这 4 个要素对新闻价值的评价是有用的，对这 4 个特征进行抽取后，构建微博事件信息库和训练集、测试集，从而进行新闻价值模型的计算，这个计算过程中也需要进行模型的计算和调整。以交通事故为例，通过这个过程可将某一天与所有交通有关的微博信息、新闻提取出来，并对其价值进行评分，在评分过程中，

新闻事件的排名越靠前价值越高。对新闻记者而言，现在只需要看 1000 条微博便可以筛选出新闻报道的线索，不仅减轻了工作量，而且能更好地评价微博数据中可能存在的新闻线索。

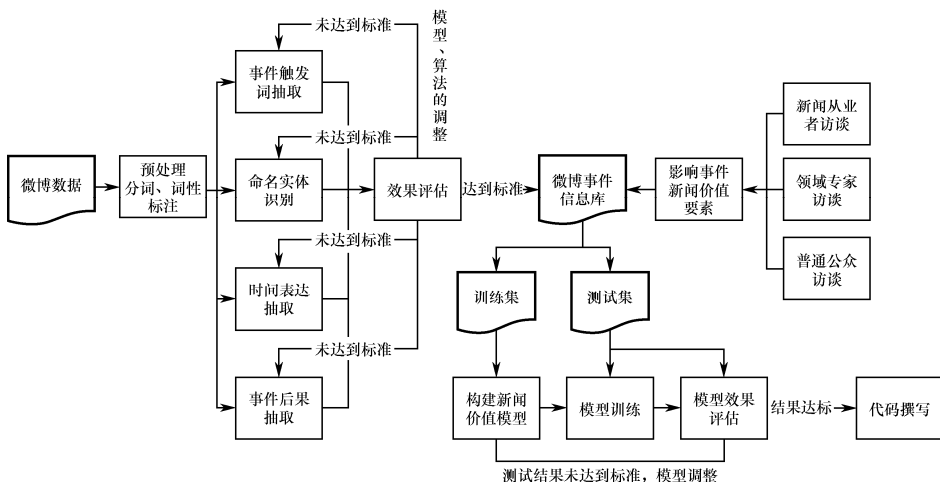


图6 技术路线图

以上便是 5 个方面的案例，其实数据是一个方面，经典管理模型的应用是第二个方面，将模型和数据结合起来，可以判断需要哪些数据、数据说明了哪些问题，以及这些数据分析如何应用到管理决策之中。

作者简介

傅湘玲：毕业于北京大学，获得管理学博士学位，现任北京邮电大学软件学院副教授，北京邮电大学社会化网络信息管理与服务中心副主任；全国信息技术标准化技术委员会 SOA 标准工作组专家成员。主要研究方向：社交网络分析。

大数据产业中的协同创新 ——技术、应用与新业态的区域实践

厦门大学自动化系副教授 洪文兴

软件和信息服务业是厦门市着力打造的千亿元产业链之一。厦门市的相关软件信息企业，对新兴技术抱有浓厚的兴趣，产生了一批耳熟能详的上市企业。这些上市企业，如美亚柏科，大多和大数据有关系。

一、概述

（一）关于姓氏的大数据

（1）数据：一组脱敏（脱敏：把敏感数据或信息去掉）后的公民个人数据，覆盖 1000 多万人，大概占全国人口的 1%，信息包括姓氏、年龄、籍贯城市、现居城市等。

（2）模型：定义一些数学模型来解释上述数据，具体定义了以下 4 个数（也是所谓的 4 个模型）：

普遍指数——姓氏人口数量。

抱团指数——各姓氏在各省分布的基尼系数。

奔波指数——姓氏人均迁徙距离。迁徙距离为户籍所在地市级与现居住地所在地级市的距离。

风雅指数——名字中不包括“取名用字频率 TOP1000”的人数占该姓氏总人数的比例。

（3）专家知识：对规律（模型反映的结果）进行解读、完善模型（所谓的专家是对一个行业比较了解的人）

因此，大数据研究至少需要涵盖数据、模型、专家知识 3 个方面。

（二）大数据基本态势

2013 年被称为“大数据元年”，经过 5 年的发展，大数据目前已进入第二个阶段，人们将其称为“大数据 2.0”。以下是对大数据 5 个方面的观察。

（1）大数据意识润物无声。大数据意识已经普遍被认可。

（2）大数据技术蓬勃发展。

（3）大数据应用随处可见。

（4）大数据产业布局加速。

（5）大数据科学呼之欲出。截至 2017 年，教育部批复了 35 所高校设置数据科学与大数据技术专业。

（三）厦门市的大数据产业规划

厦门市于 2015 年率先发布“大数据产业应用和发展规划”。该规划被称为“5+5+13 规划”，包括 5 大任务、5 大工程、13 个项目。事实上，各城市的规划都是结合自身的经济、社会和产业发展情况来编制的。厦门市的 5 个任务分别如下：推进政府大数据的开放和价值开发、推动大数据与产城融合的示范应用、加强技术创新抢占生态系统制高点、培育和引进大数据生态链产业集群、完善公共服务平台和产业发展环境。

厦门市希望将政府的大数据进行开放和开发，因此，提出了政府大数据融合共享工程，包括对外的数据门户、对内的政务信息共享协同平台。民生方面需要在厦门比较擅长的一些领域展开，如交通、教育、医疗、气象等。另外，厦门市在产业转型升级、社会治理和公共服务方面均有相关的工程。

厦门市促进大数据发展的工作实施方案包括夯实大数据发展承载基础、构建政务数据共享体系、运用大数据提升政府治理水平、运用大数据提升公共服务能力、运用大数据推动相关产业发展、强化大数据应用安全管理。

厦门市大数据工作重点包括成立大数据专家咨询委员会、制定《厦门市政府大数据开放暂行管理办法》、建设医疗健康大数据中心、建成交通大数据分析应用平台、建设统一的城市公共安全管理平台。

介绍厦门市的例子是为了说明大数据在全国乃至全世界已经进入一个新

的发展阶段。当然，虽然数据是无国界、全球化的，但还是要注意区域生态的差别。

二、技术

我们以中国计算机学会大数据专家委员会（简称 CCF 大专委）每年发布的《大数据发展趋势预测》来说明这个问题。

从 2013 年“大数据元年”开始，关于数据科学与大数据的论述还是比较粗浅的，都是一些最初的结构化论述。2016 年以后，很多问题越来越具体，越来越多样化，越来越接近大数据应用本身的特点。例如，2016 年提到了大数据的平民化、《促进大数据发展行动纲要》的推广，2017 年则希望大数据在技术上有所突破，政策法规为整个数据科学与大数据产业保驾护航。例如，在 2017 年，机器学习继续成为智能分析的核心技术，多学科融合发展，数据科学也兴起了。因此，整个学科体系、研究路线越来越清晰，国内关于大数据的学术研讨会也越来越多。

一个普遍的观点认为，大数据技术应该在企业中再次得到突破，原因是大数据所需要的数据、计算能力、专家知识等在企业中优势比较明显。

三、应用

在 CCF 大专委组织的预测中，大数据应用在哪一个领域最靠谱、最值得推广呢？答案从 2013 年至今还未发生变化，据统计，大数据应用在互联网与电子商务、金融、健康医疗 3 个领域最为广泛。

实例 1：厦门人才网数据分析项目

其目标是使找工作的人找到合适的职位、公司找到合适的求职者。求职的一般方法是到人才网搜索，而这里的方法是为人才和职位进行建模。为人才建模的问题在大数据领域被称为用户画像。用户画像可以形象化地画一个人的样子，将人的特征画在人的剪影上。简单而言，便是将人的特征分门别类地表达出来，将一个虚拟的人与现实的人对照起来。通过用户画像，能对单个人、一群人或一批人进行分析。这里涉及对人的简历的处理，包括一些

非结构化的文本数据、音频、视频数据，这些数据被称为异构数据。对异构数据的处理是大数据领域中一项很重要的技术。

实例 2：网站访问日志处理

这个项目是我们在校内做的一个新闻阅读网站，名为“一起读”。学生在访问网站的同时会留下网站访问日志。这些日志每秒都会留存网页访问请求，记录相应的 IT 来源、身份信息、访问内容、访问时间等。

我们对读者进行用户画像；新闻被称为内容，也可以进行内容画像。这样既有了内容特征，也有了用户特征，接着对新闻和用户之间的关联性进行分析。今日头条的新闻推荐也是上述工作机理。

实例 3：对专家进行讨论

所谓专家，是指在某一方面、某一领域具有比较好的知识体系的专业人士。如何刻画专家在相关领域的知识表达？最简便的方法是将该专家发表的论文、工作、专利、著作、演讲及相关新闻找出来，进行数据抽取，这样便可对专家进行画像，形成一个虚拟专家，用这个虚拟专家来表达现实世界中具备的能力。另外一个场景如工厂中遇到的技术难题，需要请专家将这个难题表达清楚，然后围绕专家的知识表达寻找解决方案。

上述 3 个例子都是通过互联网进行的。这里涉及两组对象，一组是人，另一组是物，大数据的工作就是分析人和物之间应该怎么做。

一个好的大数据应用应该包含以下 3 个要素：①专家知识，对某个场景进行解释；②数据，主要通过云计算进行保障；③数据科学技术，对其进行统计学、数据建模等角度的分析。以上 3 个部分都满足，才可以称得上一个好的大数据应用。

四、业态

图 1 所示为 2016 年全球大数据产业全景图。

中国各个城市都在画各自的产业地图，通过产业地图可以发现哪些企业在搞大数据，大数据包括哪些环节。

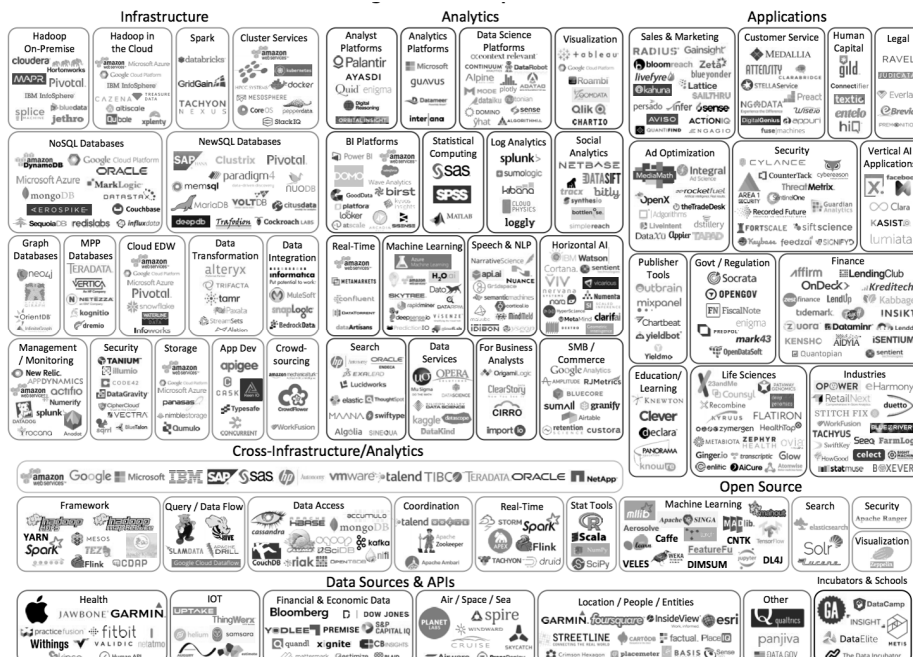


图1 2016年全球大数据产业全景图

五、实践

2016年，国家开始启动健康医疗大数据试点工程，厦门市成为第一批入选试点城市之一。面对的问题是如何将数据、科研、平台、产业、应用结合起来，形成一个闭环，形成一个区域小生态。对数据的分析交给科研单位，科研单位分析完后形成了一些共性的技术平台，这些平台可以帮助产业或企业做很多应用，这些应用又会产生很多数据，因此，一个基于大数据的完整的闭环便形成了。对高校而言，最大的产出主要是在科研（技术水平突破）和人才两个方面。

2016年开始，厦门开始举办年度国际大数据大赛，邀请了上述闭环的各个角色积极参与，是对区域产业生态的综合考验。

关于厦门信息产业和信息化研究院（厦门大学）

厦门信研院是厦门市经信局和厦门大学共建的一个产业研究平台，当前主要的研究领域是大数据、人工智能及新一代信息技术。最大特点在于研究

院的研究与产业非常接近，立足于地方产业发展，面向国家重大战略需求，研讨产业生态的本地化实现。

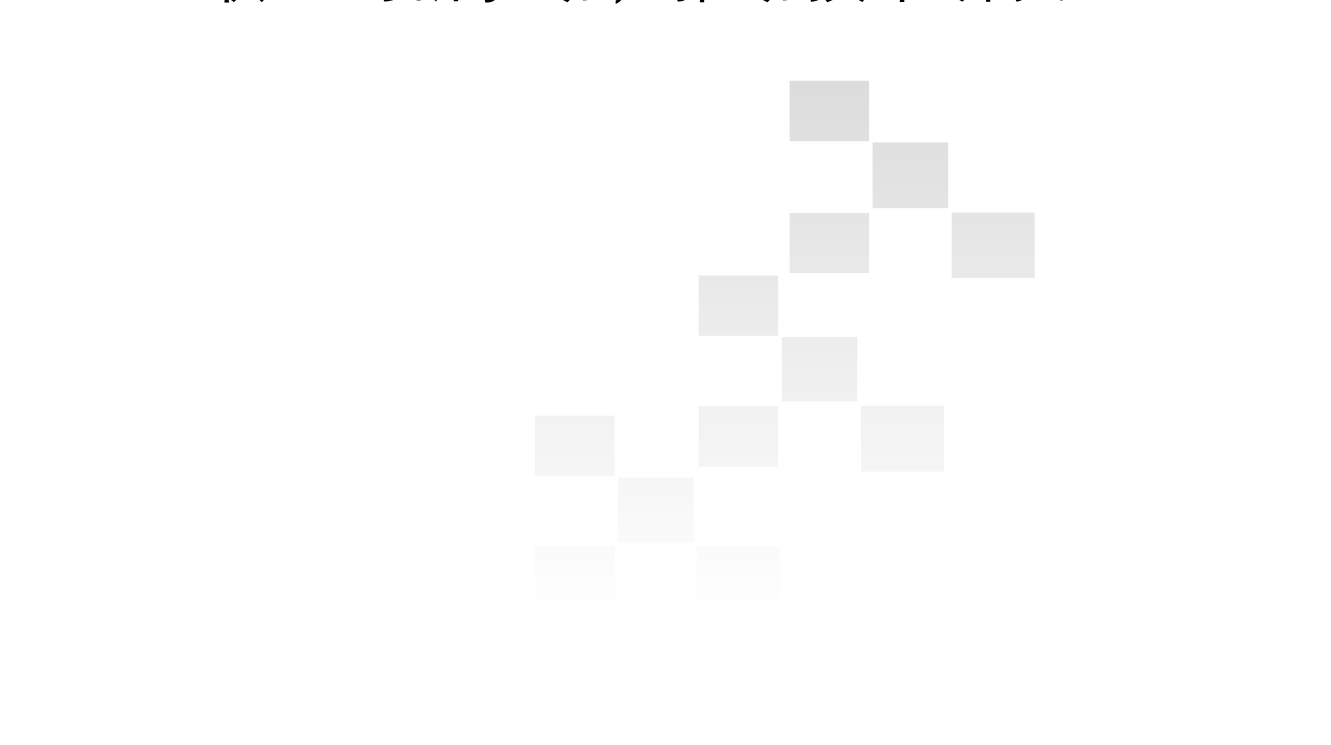
作者简介

洪文兴：厦门大学自动化系副教授，厦门信息产业与信息化研究院执行院长。2010 年获得工学博士学位，为国家公派（厦门大学与新加坡南洋理工大学）联合培养博士。2013 年起，先后任厦门大学信息科学与技术学院院长助理、厦门市信息化局软件园服务处副处长、厦门大学发展规划办公室副主任，中国系统工程学会监事、福建省系统工程学会秘书长、中国计算机学会 YOCSEF 厦门分论坛副主席（2015—2016 年）。主要研究方向为数据挖掘、大数据分析、推荐系统、软件和信息服务业产业研究、系统工程。



技 术 篇

互联网创新驱动，推动技术研发



大数据技术与产业中的几个关键问题商榷

太原科技大学计算机科学与技术学院院长 陈立潮

大数据技术和大数据产业已成为人们热议的话题，整个社会已感受到大数据对未来社会发展带来的作用与价值，掌握和应用大数据的能力已成为国家竞争力的重要体现。因此，各个国家纷纷将大数据作为国家发展战略，并给予了足够的重视。

近年来，尽管人们都在积极参与和开展大数据技术的应用研究，发表了不少关于大数据技术的报告，但却忽视了大数据技术产业的基本问题。因此，下面结合实际对大数据技术产业中需要关注和研究的若干关键问题进行分析。

一、数据资源的收集与开放

习近平总书记指出：“建设全国一体化的国家大数据中心。”显然，建设一流的大数据中心是我们开展大数据技术的关键所在。如果没有很好的大数据，便不可能进行大数据的分析与挖掘。那么，如何采集与收割分散在各个系统与环节中的数据？采取什么原则和规律对数据进行合理的采集与收割？我们不能盲目地为了收集数据而收集，需考虑到为什么收集，收集什么数据，以什么方式收集数据，以及如何合理地收集数据等问题，更不要在收集数据的过程中破坏了数据的原生态。好比一种果实，在其还未成熟的情况下便进行收割，这样的数据不仅没有任何利用价值，甚至会产生不良的结果。

在我们的信息系统中，数据是无所不在的，问题在于如何合理地收割那些有用的数据为我们服务。不要追求数据的大而全，不要看见数据便盲目地全部收割，要合理、适时、适当地收集，否则，会收集到一些无用的数据，进而影响数据挖掘的效率和效果。我们知道，错误的数据往往会导致一些错误的结果。

在数据的收集过程中，由于原生态数据的格式规范、数据的表现方式存在一定的差异，因此，收集到的数据需进行适当的预处理，要及时对数据进行一定的标注和说明，建立统一的国家级数据存储标准与规范，这也关系到以后对数据的共享问题。否则，收集到的数据将无法被很好地利用。

美国公布的一份长达 35 页的《2016—2045 年新兴科技趋势报告》中指出，在 2015 年，人类总共创造了 4.4ZB（相当于 44 亿 TB）的数据，这个数据大约每两年就会翻倍。在这些数据中隐藏了各种关于消费习惯、公共健康、全球气候变化，以及其他经济、社会、政治等方面的深层次数据和信息。可惜的是，虽然“大数据”成了一个热点，但每年只有不到 10% 的数据会被分析和利用。这不能不对数据质量和数据收集方式产生质疑。

目前，我们所收集的数据资源质量不高，数据资源流通不畅，数据价值难以被有效地挖掘和利用。因此，需要全面提升我国的大数据掌控能力、数据收集开放与共享的理念，完善相关制度，推动数据资源开放和信息流通。

二、数据资源的整合与共享

大数据的本质是怎样去发现数据中的价值，大数据的魅力在于对未来发展的预知。人类思维中所包含的信息量是相当广泛的，涉及各种各样的数据和信息。大数据中也包含各种各样的结构化与非结构化数据，且非结构化数据占据着主导地位。因此，如何将各种类型的信息进行有机整合是大数据挖掘中的一个关键问题。

大数据整合的目的有两个：一是为了降低数据再利用的成本；二是为了提升数据的非可视价值。如果将各部门的业务数据进行整合，定会提升服务对象全面而精确的数据价值。

当对不同的数据进行收集、整合后，从中会发现不同的信息与知识。对于整合起来的不同数据，从不同角度去分析，也会得到不同的结果。换言之，当我们获得一些数据后，可通过采用不同的挖掘算法发现数据中不同的有用信息和知识；当数据信息挖掘算法确定后，对于不同的数据组合也会有不同的结果。往往我们想得到的知识便抽象地隐藏于这些大数据之中。

数据的整合类似于化学反应。当我们将不同的数据整合到一起后，会产

生不同的信息与知识。问题在于如何整合或融合这些数据，如何发掘其中的信息与知识，这便涉及如何合理、有效地对大数据进行预处理。对可相融的数据进行合理的整合，进而产生新的、有价值的数据集合。

一直以来，大数据的价值属于交易过程中最棘手的问题。如果我们为数据资源定价，所交易的数据价值如何衡量？若仅仅以数量来衡量数据价值，那么大数据技术便失去了其本身的意义。

其实，数据的价值一方面来源于数据自身所带来的价值（直接价值或显式价值），另一方面是数据从其他数据的集合所产生的价值（间接价值或隐式价值）。因此，数据的收集固然很重要，但数据的整合利用更重要。如果忽视了这一点，则失去了大数据技术的魅力。

三、图像处理是未来大数据的主战场

过去我们所收集到的 44 亿 TB 数据中，大多属于非结构化数据，特别是一些视频图像数据。随着近年来智慧城市在我国的普及与推广，视频图像的收集成为大数据的主要来源，因此，图像大数据处理与分析也成为了大数据技术的研究热点。

我们知道，人类思维活动的大量信息主要是通过眼睛来获取的，人的一生中的知识来源也是用眼睛来承担的，占有 80% 的比例。图像分析与处理便成为大数据技术的应用基础，也是未来人类社会、人类智慧和人工智能的突破点。

人脸识别技术便是图像识别处理中的热点问题，它是智慧城市建设过程中的一项关键技术，不仅是银行、医疗、保险行业身份的识别与辨认，对于城市的治安、人类的活动轨迹等均可提供相应的决策支持。

图像大数据的一个很大的特点在于其连续性，需要采取一种流式处理的方式，这与我们以往考虑的静态图像不同。因此，图像处理是一种动态数据，连续、迭代分析是图像大数据处理的难点问题。图像大数据处理的另一个关键问题是要建立相应的一些图像资源库，这也是一项基础性工作，也是图像处理的基础。

四、从数据可视化到数据透视化

数据可视化是大数据分析的一种基本方式，它可以将一些繁杂凌乱的数据以各种可视化的方式（如图、表等）呈现出数据的规律与变化特点，但这种可视化仅仅反映了数据的表象特征。由于我们所涉及的数据已远远超出了平面思维的状态，如果从不同角度看待这些数据的本质，则需另一种方式，即透视化技术。

数据的透视化主要是从数据的多维角度来观察数据的形态，从多维空间角度来刻画和理解大数据中所包含的深层次信息与知识，属于数据可视化的一直扩展和延伸。数据透视化的关键主要是如何确定数据处理与分析中的透视点和视角，正如立方体图形的视点不断改变时，其所呈现的图形效果是不同的。对一个立方体数据而言，当我们看待立体的角度不同时，大数据会带来不同的结果。

过去我们所考虑的数据大多为数据本身的价值，如今的数据越来越离不开时间和空间的约束，如果抛开时间和空间的概念单独分析数据，数据的作用和意义便不会太大了。例如，股票数据具有时间和空间的特征，如果不考虑时间和空间特征，股票数据便失去了意义。

因此，透视化技术的另一个含义就像 X 射线那样，能否通过一种机制和算法来探索数据中可能隐含的信息价值，这也许是一种不可能的设想。但随着挖掘算法和大数据技术及人工智能技术的发展，也许能找到实现这种技术的可能。

我国大数据产业“十三五”规划指出，加快发展面向大数据分析的在线机器学习、自然语言处理、图像理解、语音识别、空间分析、大数据可视化等数据服务，在这些数据分析服务中无不涉及时间和空间的问题，使数据分析变成了多维数据处理问题。

五、智慧城市中的大数据

智慧城市是基于数字城市、物联网和云计算建立的现实世界与数字世界之间的融合，以实现对人和物的感知、控制和智能服务，智慧城市的实现需

要建设更加完善的信息技术设施，以及包括智慧城市运营为主的技术支持。智慧城市建设中所产生的大数据是推动智慧城市发展的原动力，需要有针对性地加快大数据的技术创新和重点攻关研究，这样才能推动和加速智慧服务产业的发展。

大数据是智慧城市中各领域均可实现智慧化的关键性支撑技术。智慧城市的建设离不开大数据，大数据也会遍布在智慧城市的各个方面。从政府的决策与服务到人民的衣食住行生活方式；从创建节约型社会到以人为本；从科技惠民再到城市的产业布局 and 规划，指导城市的运营和管理等，都在大数据的支撑下走向“智慧化”。

然而，智慧城市的兴起导致城市数据中心的急剧增加，许多城市为了建设智慧城市而盲目投资城市数据中心，在一个城市建立多个数据中心，造成了数据存储分散、数据中心过剩、数据资源浪费、能源损耗过量等现象。

目前，我国的智慧城市信息化建设管理平台还不成熟，数据标准和规范还需进一步健全，城市数据中心的基础管理系统解决方案存在模式多样、功能不一、架构自主，造成了城市公共信息管理混乱的问题，还出现了新的数据孤岛，极大地阻碍了新型智慧城市的建设步伐。

其实，城市智慧存在于城市的运行机制之中，存在于城市各方面资源的配合与协作之中，仅仅依靠城市数据中心的简单建设是解决不了问题的。智慧城市建设的过程中，最关键的问题在于各种数据资源的系统整合、综合分析 with 智能决策，如何利用大数据使城市的运行智慧化是智慧城市建设的核心问题。因此，如果无法有效地利用大数据技术，智慧城市的基础设施也会沦为监控系统。可以说，大数据技术正是智慧城市建设和运营的基石。

数据世界并不能代替真实世界，在城市的建设过程中，隐含着城市的文化、经济、民俗等信息，也包含着城市未来的趋势、社会层级、人类欲望等，这些都无法很好地用数据来准确表达。智慧城市的大数据技术也仅仅是对城市运行的宏观决策。因此，过分依赖大数据技术，也会影响决策的准确性、社会性和真实性。

总之，大数据技术为城市规划带来了巨大的影响，推动了智慧城市发展。智慧城市发展使整个城市更加便捷、科学、合理地规划，大数据也将极大地提高政府部门的决策效率和服务水平。

目前,大数据技术应用进展缓慢的瓶颈恰恰是一些最基本的问题没有得到很好的解决。不管是大数据技术与产业中的关键问题,还是对大数据技术未来发展的一些设想,但愿这些不成熟的甚至是幼稚的观点能对未来大数据技术起到抛砖引玉的作用。

作者简介

陈立潮:山西万荣人。2003年3月毕业于北京理工大学计算机应用技术专业,获工学博士学位。多年来一直从事于高校教学与科研工作,主要研究方向为软件工程、数据挖掘、模式识别与图像处理等。担任教育部高等学校大学计算机课程教学指导委员会委员、全国高等学校计算机教育研究会常务理事、中国仿真学会理事、全国高等学校大数据联盟理事、中国计算机学会软件工程专委会委员、中国高校计算机教育MOOC联盟山西工作委员会主任、中共山西省委高级联系专家、山西省计算机学会副理事长兼秘书长、山西省软件行业协会常务理事等职务。

Python 编程要点

山东工商学院计算机科学与技术学院副教授 董付国

一、Python 语言的特点和优势

（一）免费开源跨平台

免费开源是备受人们喜爱的，跨平台也备受人们喜爱。Windows 平台、各种版本的 Linux、苹果系统、Android 手机均可使用 Python 语言。

（二）易学易用

与 C 语言、C++、Java 等语言相比，Python 语言更容易上手，几天便可入门，一两个月之内便可写出很好的程序。

（三）简洁清晰，可读性强

代码简洁清晰，赏心悦目。一个好的 Python 代码不仅是正确的，还应该是漂亮的、优雅的。

（四）功能强大

本身有大量运算符、内置对象和标准库对象。

以图 1 所示的代码为例对代码进行解释。第一行是导入 random 模块，是一个标准库；第二行的 range(20)是返回的一个 range 对象，range 对象包含了从 0 到 20 的数字，包括 0 但不包括 20，是一个左闭右开的区间；第三行的 random.shuffle(x)，shuffle 是随机打乱顺序，像扑克牌洗牌一样。接着对打乱的数字进行排序，在 Python 中直接用 x.sort()对数字进行排序，其中可以为 sort 加一个参数，实现降序排列，制定排序规则，实现更复杂的排序功能。

```
...
>>> import random
>>> x = list(range(20))
>>> random.shuffle(x)
>>> x
[6, 2, 4, 13, 7, 18, 10, 12, 15, 16, 19, 3, 5, 8, 11, 0, 14, 1, 17, 9]
>>> x.sort()
>>> x
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]
>>>
```

图 1 代码示例

以图 2 为例解释 Python 语法的简洁清晰。第一行的意思为交换两个变量的值，它属于 Python 中的序列解包用法。一般而言，将两个变量的值进行交换，需借助于一个中间变量，如 $c=a$, $a=b$, $b=c$ 。但在 Python 语言中不需要中间变量， $a, b=b, a$ 便可；第二行中相当于数学中的不等式，在 Python 中的关系运算符可以连着用， $1<2<3<4$ 等价于 $1<2$ and $2<3$ and $3<4$ ，可省掉很多字符；第三行是一个列表，其中有 1, 2, 3 的元素，用 “in” 关键字来查看 3 是否在列表中，在列表中时返回 “true”，不在列表中时返回 “false”；第四行是导入 randint，接着使用列表推导式生成 10 个 1 到 1000 的数字。

```
>>> a, b = b, a
>>> 1 < 2 < 3 < 4
True
>>> 3 in [1, 2, 3]
True
>>> from random import randint
>>> lst = [randint(1, 1000) for i in range(10)]
```

图 2 Python 语法

图 3 所说明的问题是 Python 对于代码布局的要求非常严格，可读性强，要求程序不仅要正确，还要漂亮。如代码的缩进，来体现代码的业务逻辑关系，如果缩进不正确，意味着程序是错误的。

图 4 为 Python 所提供的运算符，比其他语言丰富，且每一个运算符的功能也比其他语言强大。如 “+” 不仅可用作算术的加法运算，还可将两个列表、元素或字符串连接起来。“-” 不仅可用作算术减法，还可用作集合差集、相反数等。其中，隐含着一个知识，即 Python 内置支持集合及各种运算，并且支持复数及加、减、乘、除等各种运算。

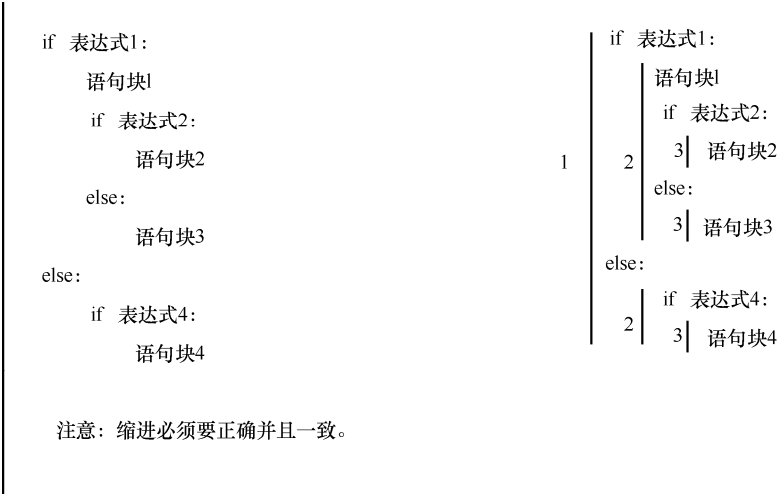


图 3 Python 对于代码布局的要求

运算符	功能说明
+	算术加法，列表、元组、字符串合并与连接，正号
-	算术减法，集合差集，相反数
*	算术乘法，序列重复
/	真除法
//	求整商，但如果操作数中有实数的话，结果为实数形式的整数
%	求余数，字符串格式化
**	幂运算
<, <=, >, >=, ==, !=	(值) 大小比较，集合的包含关系比较
or	逻辑或
and	逻辑与
not	逻辑非
in	成员测试
is	对象同一性测试，即测试是否为同一个对象或内存地址是否相同
~, ^, &, <<, >>, ~	位或、位异或、位与、左位移、右位移、位求反
&, , ^	集合交集、并集、对称差集
@	矩阵相乘运算符

图 4 运算符

(五) 生态良好

Python 有大量涉及各领域的专业扩展库，有很多狂热的 Python 支持者与热爱者开发出很多可用的扩展库。

图 5 演示了 Python 扩展库 matplotlib 中 pyplot 可视化和扩展库 sklearn 中系统聚类算法 AgglomerativeClustering 的应用，生成的效果图如图 6 所示。

```
def AgglomerativeTest(n_clusters):
    '''聚类，指定类的数量，并绘制图形'''
    assert 1 <= n_clusters <= 4
    predictResult = AgglomerativeClustering(n_clusters=n_clusters,
                                            affinity='euclidean',
                                            linkage='ward').fit_predict(data)

    colors = 'rgby'
    markers = 'o*^v+'
    for i in range(n_clusters):
        subData = data[predictResult==i]
        plt.scatter(subData[:,0], subData[:,1], c=colors[i], marker=markers[i], s=40)
    plt.show()

# 生成随机数据
data = generateData()
# 聚类为3个不同的类
AgglomerativeTest(3)
# 聚类为4个不同的类
AgglomerativeTest(4)
```

图5 Python 扩展库中聚类算法的应用

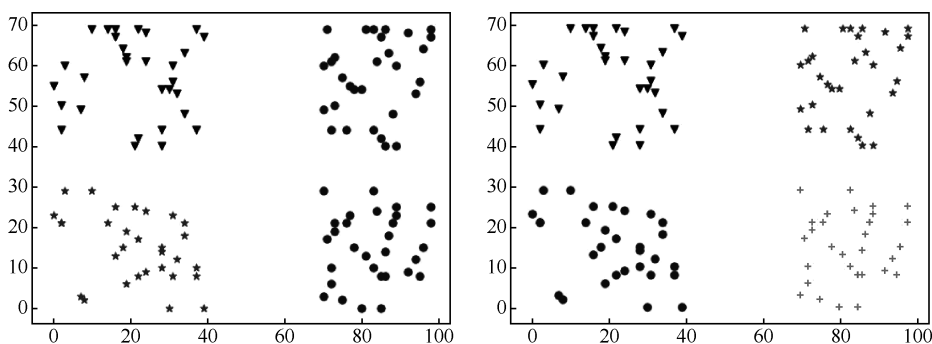


图6 聚类算法应用生成效果图

二、Python 所支持的内置类型

int、float、complex 分别是数字、整数、实数和复数类型；str、bytes、ascii 是字符串和字节串类型；接着是常用的数据结构，如列表、元素、集合及冻结的集合；range、map、zip 等是对象及函数，除第一个外，其他都具有惰性求值的特点。如 map 函数返回的是 map 对象，其中访问过的元素无法再次访问，zip、filter 等对象也具有同样的特定，不支持随机访问，不支持使用下标访问其中的任意一个元素。

Python 中可支持非常大的数字，如 99999 的 99 次方，可直接算出结果，无任何障碍。

图 7 简单列出了 Python 常用的扩展库，第一行与数字、统计、随机有关；第二行与字符串及表达式有关；第三行与系统编程有关，当然在系统运维中也会用到。

```
• math、decimal、fractions、statistics、random
• string、re
• sys、os、os.path、shutil、platform、ctypes
• collections、heapq、queue、array、enum、bisect
• itertools、functools、operator
• datetime、time、calendar
• json、pickle、struct、shelve、marshal、sqlite3、zipfile、tarfile、gzip、csv
• threading、multiprocessing、subprocess、asyncio、concurrent
• socket、urllib、http、smtplib、ftplib、poplib、email、ssl
• tkinter、turtle
• pdb、timeit、unittest、doctest
• hashlib、zlib、hmac
```

图 7 Python 常用标准库

图 8 所示为常用的 Python 扩展库。第 1 行与图形图像有关，pillow 库是做数字图像处理常用的库，pyopengl 是封装图形学的库，pygame 可以做游戏策划、游戏处理及简单的游戏编程。第 2 行是科学计算可视化领域应用较多的，pyopencv 在计算机视觉和图像处理利用较多。第 3 行在机器学习、深度学习等领域用的较多，尤其是 tensorflow 在深度学习领域应用非常广泛，pySpark 在大数据领域应用比较广泛。第 4 行是在密码学领域应用非常广泛。第 5 行是在网页编程用的较多。第 6 行是在 GUI 开发中常用的扩展库。第 7 行在自然语言处理、中英文分词中用的较多。第 8 行在系统运维中用的较多。第 9 行在网络爬虫中用的较多。第 10 行是在数据库访问领域用的较多。第 11 行是在软件安全领域用的较多。最后一行用于把 Python 程序打包成可执行文件。

- pillow、pyopengl、pygame
- numpy、scipy、matplotlib、Bokeh、VisPy、Glumpy、Seaborn、NetworkX、pyopencv、pandas
- pycuda、pyopencl、theano、scikit-learn、tensorflow、NumbaPro、pySpark、Keras、PyBrain、Milk、Orange、XGBoost
- pycrypto、rsa
- django、flask、web2py、Pyramid、Bottle
- wxPython、kivy、PyQt、PyGtk、Page for Python
- jieba、snownlp、pypinyin、chardet、NLTK、Pattern
- xlrd、xlwt、openpyxl、python-docx、python-pptx、pdfminer3k
- psutil、pywin32、scapy
- scrapy、BeautifulSoup4
- pymssql、pyodbc、MySQLdb、pymongo、cx_Oracle
- idaPython、Immunity Debugger、Paimei、ropper
- py2exe、pyinstaller、cx_Freeze

图 8 Python 生态圈

Python 的内存管理与其他语言不同，它是基于值的自动内存管理方式，这种方式在变量中并不直接存放值，而是存放值的引用。而 C 语言中每个变量是直接存储值的。

在 Python 中编程，不需要提前声明一个变量，如果需要一个变量，可以直接对其赋值，会自动创建变量。要注意的是，虽然不需要说明每一个变量的类型，但在每一个瞬间，每一个变量均属于一个特定的类型。

在列表对象的使用过程中需谨慎，列表对象功能虽然很强大，但负担也很重，开销很大。使用过程中应尽量避免在列表非尾部位置进行元素添加或删除操作。列表的 in 操作的时间复杂度为 $O(n)$ ，而集合和字典的 in 操作是常数级的。

一般而言，操作一个文件分三步走：第一步，打开文件；第二步，操作文件；第三步，关闭文件。操作文件包括读、写、修改等，在操作文件的过程中需特别注意确定文件是否关闭，由于代码问题、磁盘空间不足、网络突然中断等问题会出现异常。如果使用上下文管理语句 with，则可有效避免这个问题。

在解决问题时，可能会有很多种实现的方法。首先要考虑算法级别的优化，以及语言机制的优化，如选择不同的数据类型，可能会对程序的执行速度产生影响。

在使用 Python 编程时，可考虑使用一些机制，在机制层面上对代码进行

优化，如 `functools` 库中的 `lru_cache`。

在写代码时，要养成写注释的习惯，以方便日后查看代码，回顾写代码的思路。也可方便他人调用函数时显示其使用帮助，查看其调用形式及功能。

作者简介

董付国：副教授，先后出版《Python 程序设计》《Python 程序设计基础》《Python 程序设计（第2版）》《Python 可以这样学》《Python 程序设计开发宝典》《中学生可以这样学 Python》《Python 程序设计基础（第2版）》等系列教材，总印刷量超过4万册，全国近100所学校选作教材，国家图书馆和国内大部分学校图书馆购买了系列图书供读者借阅，其中《Python 可以这样学》已被引入中国台湾发行繁体版。

2016年应邀在“第6届高等学校计算机程序设计课程研讨会”上做题为“Python 可以这样学，Python 应该这样教”的大会报告。近两年担任超过10期全国高校/中学教师 Python 编程与应用培训班的主讲教师。2017年10月应邀担任“全国高校 Python 课程高级研修班”3位主讲教师之一。2017年11月入选第二届“中国大数据创新百人榜单”，获“中国大数据学术创新奖”。

长期维护微信公众号“Python 小屋”并免费分享450多篇 Python 技术文章，关注人数超过1.3万，访问量超过60万。

在学术中国、CSDN 学院、龙果学院等平台开设4次直播课和8套视频录播课，总数超过600课，受众人数超过4万。

数据可视化

北京邮电大学世纪学院通信与信息工程系副主任 刘 刚

一、数据可视化

就大数据而言，其本身是一座矿藏。如稀土矿，表面上看仅仅是一堆土或岩石，很难看出数据的价值。因此，大数据的可视化包含两方面内容：数据的挖掘和数据的有效呈现。

从总体数据可视化而言，应注意以下两点：从用户出发、从客观科学出发。

从用户出发，需使数据变得形象易懂，使其在阅读时感受到舒适、颜色配比合适，在可视化呈现中，迅速获取重点内容，最后一个也是最重要的，即为用户呈现的内容是真实可靠的。

二、理论与应用研究

为做到以上几点，便需要在数据分析时遵循科学性原则。

如何做到科学性？要在模型设计、数据筛选、数据分析和数据挖掘几方面下功夫。因此，基于这几点，笔者将若干个项目（其中有国家级层面的，也有企业应用需求的）做了以下总结：

（一）模型调研工作

将国内外所有的可视化研究进展成果进行调研，如 Google、百度的可视化，调研其可视化会用到的工具。

（二）可视化模型调研

包括对美国国家统计局、人口局、世界银行、英法德等先进国家的具有政府大数据的可视化呈现模型的调研，另外，还对典型的可视化用到的统计

分析模型进行了调研，以及对数据模型、挖掘模型的调研，这方面形成了 8 个调研报告。

基于以上工作基础做了一些工程项目，将工程项目中的政府或大型企业所具有的信息数据以数据图、模型库的形式装入到自己开发的图库工具之中，完成大数据可视化的呈现体系。简单的数据呈现用 Excel 也可以实现，利用一些简单的饼图、柱状图、折线图进行分析。但对于真正的大数据而言，由于数据的属性、维度很多，如空间属性、时间属性、地理属性等，以及一些行业的分类属性要求，简单的饼图、折线图很难满足大数据的可视化要求。

三、数据分析模型

（一）数据的基本呈现

数据分析的模型有很多种，首先，如果要准确地掌握我们所了解的数据及数据模型和数据分析间的对应。如果数据模型不正确，得到的数据可信度便会丧失。有了数据模型后，需将数据进行图形化的展示，具体应做以下几方面的工作：第一，要关联数据，将模型和数据做好关联；第二，对数据进行层级分类，分析数据具体属于哪个层次、维度；第三，对数据维度的处理，目前看到的数据大部分是二维数据，对于二维数据的呈现是横、竖两个坐标，用折线图、柱状图便可以表示。二维数据的呈现形式是较为单调的，所表达的寓意不够丰富，很难将多个指标间的内在关系进行表达。因此，如何对数据进行维度的表示也很重要。做好这些工作后，便可以分析数据坐标的生成。

1. 数据表述关系

首先以二维数据为例，分析关联数据如何表达表述关系。在做表述时，可以利用流程图、网络图或表格图的形式将数据间的关系关联起来。

接着可以做一些数据对比图，用作数据的对比分析和呈现。例如，柱状图的应用，也可将柱状图画在两侧进行对比。数据类型的对比图还可以利用饼图的变异——南丁格尔图，对比图也可以通过柱状图的高矮、饼图的面积大小及柱状图的占比面积，对图形数据大小和占比进行一目了然的对比，这些用传统的 Excel 方式实现起来是比较困难的，但这些并不是对图的最复杂的表示，仅仅是两个维度或三个维度的表达。

2. 数据层级关系

在进行数据表达时，尤其是一些复杂的数据，需先对数据进行分层和分类，判断其属于哪个层级。进行层级分类表达时，也运用了很多分类的技术，这些也需有合适的算法保证对数据进行分层分类。

3. 数据信息表达

数据信息的表达与具体的数据属性及数据算法相关联，可用柱状图、面积图、动态散点图的形式来表达各个不同数据在表格中占据的位置，如在家庭消费中哪些项目占比大。

（二）数据和属性的结合

很多数据都是具有属性的，如地理属性、时间属性、空间属性等。

以地理属性为例，许多图（如某个地区的工业聚集度、人口密度、环境污染度、人口迁移等）与地理有关，可以在 GIS 地图上利用色彩的明亮或高亮等形式将数据的大小分布在地图上。因此，在进行数据展示时，也可以使用 GIS 形式作为数据的入口，如文化产业法人单位的统计，中东部较多，西部、北部较少等。

对图形的表述内容很多，可以将数据分为点图、线图、面图，再进行分类表达。

四、北邮 Chart 系统

为使数据可视化更好地表达，北邮也做了一套自己的系统——北邮 Chart，以更方便地表达数据可视化。在这套系统中，做了以下几方面的内容。

（1）**数据地图**。做到“一图知天下”，将与数据相关的地理信息加在地图上。

（2）**制图工具**。图形分为 31 个大类、100 多张，在图中对 20 多种参数进行优化。

（3）**数据分析工具**。主要用于数据管理、科学管理、大数据方面的研究。

（4）**专业应用**。定制个性化专业图，用于更好地表述。因为数据属性不

同，不能只用简单的柱状图、折线图、饼图表示所有的图形。

(5) 用户作品。运用北邮 Chart 系统，用户可以自己生成和保存研究成果。

对数据可视化的展示，除传统的简单图形外，还有很多复杂的表示，大数据可视化的表示不应是简单的静态表示，而是利用静态与动态相叠加的表达方法来呈现。

数据的表达其实并不容易，首先要准确理解需求，并能找到合适的可视化图形，要易于理解。当然，对某类数据的表达方式可能有多种，需要做出选择。

作者简介

刘刚：北京邮电大学副教授、硕士生导师，北京邮电大学世纪学院通信系副主任。北京邮电大学超图大数据中心副主任。主要从事数据分析和数据可视化工程研究。

大数据时代的数据挖掘

南京邮电大学计算机学院院长 李 涛

众所周知，大数据时代的大数据挖掘已成为各行各业的一大热点。

一、数据挖掘

在大数据时代，数据的产生和收集是基础，数据挖掘是关键，数据挖掘可以说是大数据最关键的也是最基本的工作。通常而言，数据挖掘也称为 Data Mining（或知识发现 Knowledge Discovery from Data），泛指从大量数据中挖掘出隐含的、先前未知的但潜在的有用信息和模式的一个工程化和系统化的过程。

不同的学者对数据挖掘有着不同的理解，个人认为，数据挖掘主要有以下 4 个特性。

（一）应用性（A Combination of Theory and Application）

数据挖掘是理论算法和应用实践的完美结合。数据挖掘源于实际生产生活中应用的需求，挖掘的数据来自具体应用，同时通过数据挖掘发现的知识又要运用到实践中去，辅助实际决策。所以，数据挖掘来自应用实践，同时也服务于应用实践，数据是根本，数据挖掘应以数据为导向，其中涉及算法的设计与开发都需考虑到实际应用的需求，对问题进行抽象和泛化，将好的算法应用于实际中，并在实际中得到检验。

（二）工程性（An Engineering Process）

数据挖掘是一个由多个步骤组成的工程化过程。数据挖掘的应用特性决定了数据挖掘不仅是算法分析和应用，而且还是一个包含数据准备和管理、

数据预处理和转换、挖掘算法开发和应用、结果展示和验证，以及知识积累和使用的完整过程。在实际应用中，典型的数据挖掘过程还是一个交互和循环的过程。

（三）集合性（A Collection of Functionalities）

数据挖掘是多种功能的集合。常用的数据挖掘功能包括数据探索分析、关联规则挖掘、时间序列模式挖掘、分类预测、聚类分析、异常检测、数据可视化和链接分析等。一个具体的应用案例往往涉及多个不同的功能。不同的功能通常有不同的理论和技术基础，而且每一个功能都有不同的算法支撑。

（四）交叉性（An Interdisciplinary Field）

数据挖掘是一门交叉学科，它利用了来自统计分析、模式识别、机器学习、人工智能、信息检索、数据库等诸多不同领域的研究成果和学术思想。同时，一些其他领域如随机算法、信息论、可视化、分布式计算和最优化也对数据挖掘的发展起到了重要的作用。数据挖掘与这些相关领域的区别可以由前面提到的数据挖掘的 3 个特性来总结，最重要的是它更侧重于应用。

综上所述，应用性是数据挖掘的一个重要特性，是其区别于其他学科的关键，同时，其应用特性与其他特性相辅相成，这些特性在一定程度上决定了数据挖掘的研究与发展，同时，也为如何学习和掌握数据挖掘提出了指导性意见。如从研究发展来看，实际应用的需求是数据挖掘领域很多方法提出和发展的根源。从最开始的顾客交易数据分析（Market Basket Analysis）、多媒体数据挖掘（Multimedia Data Mining）、隐私保护数据挖掘（Privacy-Preserving Data Mining）到文本数据挖掘（Text Mining）和 Web 挖掘（Web Mining），再到社交媒体挖掘（Social Media Mining），都是由应用推动的。工程性和集合性决定了数据挖掘研究内容和方向的广泛性。其中，工程性使得整个研究过程中的不同步骤都属于数据挖掘的研究范畴。而集合性使得数据挖掘有多种不同的功能，而如何将多种功能联系和结合起来，从一定程度上影响了数据挖掘研究方法的发展。比如，20 世纪 90 年代中期，数据挖掘的研究主要集中在关联规则和时间序列模式的挖掘。到 20 世纪 90 年代末，研究人员开始研究基于关联规则和时间序列模式的分类算法（如 classification based on

association), 将两种不同的数据挖掘功能有机地结合起来。21 世纪初, 一个研究的热点是半监督学习 (Semi-Supervised Learning) 和半监督聚类 (Semi-Supervised Clustering), 也是将分类和聚类这两种功能有机结合起来。近年来的一些其他研究方向如子空间聚类 (Subspace Clustering) (特征抽取和聚类的结合) 和图分类 (Graph Classification) (图挖掘和分类的结合) 也是将多种功能联系和结合在一起。最后, 交叉性导致了研究思路和方法设计的多样化。

前面提到的是数据挖掘的特性对研究发展及研究方法的影响, 另外, 数据挖掘的这些特性对如何学习和掌握数据挖掘提出了指导性的意见, 对培养研究生、本科生均有一些指导意见, 如应用性在指导数据挖掘时, 应熟悉应用的业务和需求, 需求才是数据挖掘的目的, 业务和算法、技术的紧密结合非常重要, 了解业务、把握需求才能有针对性地对数据进行分析, 挖掘其价值。因此, 在实际应用中需要的是一种既懂业务又懂数据挖掘算法的人才。工程性决定了要掌握数据挖掘需有一定的工程能力, 一个好的数据挖掘人员首先是一名工程师, 有很强大的处理大规模数据和开发原型系统的能力, 这相当于在培养数据挖掘工程师时, 对数据的处理能力和编程能力很重要。集合性使得在具体应用数据挖掘时, 要做好底层不同功能和多种算法积累。交叉性决定了在学习数据挖掘时要主动了解和学习相关领域的思想和技术。

因此, 这些特性均是数据挖掘的特点, 通过这 4 个特性可总结和学习数据挖掘。

二、大数据的特征

大数据 (Big Data) 一词经常被用以描述和指代信息爆炸时代产生的海量信息。研究大数据的意义在于发现和理解信息内容及信息与信息之间的联系。研究大数据首先要厘清和了解大数据的特点及基本概念, 进而理解和认识大数据。

业界普遍认为, 大数据具有标准的“4V”特征。

- (1) **Volume (大量)**: 数据体量巨大, 从 TB 级别跃升到 PB 级别。
- (2) **Variety (多样)**: 数据类型繁多, 如网络日志、视频、图片、地理位

置信息等。

(3) **Velocity (高速)**: 处理速度快, 实时分析, 这点和传统的数据挖掘技术有着本质的不同。

(4) **Value (价值)**: 价值密度低, 蕴含有效价值高, 合理利用低密度价值的数 据并对其进行正确、准确的分析, 将会带来巨大的商业和社会价值。

上述“4V”特点描述了大数据与以往部分抽样的“小数据”的主要区别。然而, 实践是大数据的最终价值体现的唯一途径。从实际应用和大数据处理的复杂性看, 大数据还具有如下新的“4V”特点。

(1) **Variability (变化)**: 在不同的场景、不同的研究目标下, 数据的结构和意义可能会发生变化, 因此, 在实际研究中要考虑具体的上下文场景 (Context)。

(2) **Veracity (真实性)**: 获取真实、可靠的数据是保证分析结果准确、有效的前提。只有真实而准确的数据才能获取真正有意义的结果。

(3) **Volatility (波动性) /Variance (差异)**: 由于数据本身含有噪声及分析流程的不规范性, 导致采用不同的算法或不同分析过程与手段会得到不稳定的分析结果。

(4) **Visualization (可视化)**: 在大数据环境下, 通过数据可视化可以更加直观地阐释数据的意义, 帮助理解数据, 解释结果。

综上所述, 以上“8V”特征在大数据分析与数据挖掘中具有很强的指导意义。

三、大数据时代下的数据挖掘

大数据挖掘的核心和本质是应用、算法、数据和平台 4 个要素的有机结合。

数据挖掘是应用驱动的, 来源于实践, 海量数据产生于应用之中。需用具体的应用数据作为驱动, 以算法、工具和平台作为支撑, 最终将发现的知识和信息应用到实践中去, 从而提供量化的、合理的、可行的且能产生巨大价值的信息。

要想挖掘大数据中隐含的有用信息, 需设计和开发相应的数据挖掘和学习算法。算法的设计和开发需以具体的应用数据作为驱动, 同时在实际问题

中得到应用和验证，而算法的实现和应用需要高效的处理平台，这个处理平台可以解决波动性问题。高效的处理平台需要有效分析海量数据，及时对多元数据进行集成，同时有力支持数据化对算法及数据可视化的执行，并对数据分析的流程进行规范。

总之，应用、算法、数据、平台这 4 个方面相结合的思想，是对大数据时代的数据挖掘理解与认识的综合提炼，体现了大数据时代数据挖掘的本质与核心。这 4 个方面也是对相应研究方面的集成和架构，这 4 个架构具体从以下 4 个层面展开。

应用层 (Application)：关心的是数据的收集与算法验证，关键问题是理解与应用相关的语义和领域知识。

数据层 (Data)：数据的管理、存储、访问与安全，关心的是如何进行高效的数据使用。

算法层 (Algorithm)：主要是数据挖掘、机器学习、近似算法等算法的设计与实现。

平台层 (Infrastructure)：数据的访问和计算，计算平台处理分布式大规模的数据。

综上所述，数据挖掘的算法分为多个层次，在不同的层次有不同的研究内容，可以看到，目前在做数据挖掘时的主要研究方向，如利用数据融合技术预处理稀疏、异构、不确定、不完整及多来源数据；挖掘复杂动态变化的数据；测试通过局部学习和模型融合所得到的全局知识，并反馈相关信息给预处理阶段；对数据并行分布化，达到有效使用的目的。

四、大数据挖掘系统的开发

(一) 背景目标

大数据时代的来临使得数据的规模和复杂性都呈现爆炸式的增长，促使不同应用领域的数据分析人员利用数据挖掘技术对数据进行分析。在应用领域中，如医疗保健、高端制造、金融等，一个典型的数据挖掘任务往往需要复杂的子任务配置，整合多种不同类型的挖掘算法，以及在分布式计算环境中高效运行。因此，在大数据时代进行数据挖掘应用的当务之急是要开发和

建立计算平台和工具，支持应用领域的数据分析人员能够有效地执行数据分析任务。

（二）相关产品

1. 现有的数据挖掘工具

现有的数据挖掘工具有 Weka、SPSS 和 SQL Server，它们提供了友好的界面，方便用户进行分析。然而，这些工具并不适合进行大规模的数据分析，同时，在使用这些工具时用户很难添加新的算法程序。

2. 流行的数据挖掘算法库

流行的数据挖掘算法库有 Mahout、MLC++ 和 MILK，这些算法库提供了大量的数据挖掘算法。但这些算法库需要有高级编程技能才能进行任务配置和算法集成。

3. 最近出现的一些集成的数据挖掘产品

最近出现的一些集成的数据挖掘产品有 Radoop 和 BC-PDM，它们提供友好的用户界面来快速配置数据挖掘任务。但这些产品是基于 Hadoop 框架的，对非 Hadoop 算法程序的支持非常有限。没有明确地解决在多用户和多任务情况下的资源分配。

（三）FIU-Miner

为解决现有工具和产品在大数据挖掘中的局限性，我们团队开发了一个新的平台——FIU-Miner，它代表了 A Fast、Integrated 和 User-Friendly System for Data Mining in Distributed Environment。它是一个用户友好并支持在分布式环境中进行高效率计算和快速集成的数据挖掘系统。与现有数据挖掘平台相比，FIU-Miner 提供了一组新的功能，能够帮助数据分析人员方便并有效地开展各项复杂的数据挖掘任务。

与传统的数据挖掘平台相比，FIU-Miner 提供了一些新的功能，主要有以下几个方面。

（1）用户友好、人性化、快速的数据挖掘任务配置。基于“软件即服务”这一模式，FIU-Miner 隐藏了与数据分析任务无关的低端细节。通过 FIU-Miner

提供的人性化用户界面，用户可以通过将现有算法直接组装成工作流，轻松完成一个复杂数据挖掘问题的任务配置，而不需要编写任何代码。

(2) 灵活的多语言程序集成。允许用户将目前最先进的数据挖掘算法直接导入系统算法库中，以此对分析工具集合进行扩充和管理。同时，由于 FIU-Miner 能够正确地将任务分配到有合适运行环境的计算节点上，所以，对这些导入的算法没有实现语言的限制。

(3) 异构环境中有效的资源管理。FIU-Miner 支持在异构的计算环境中（包括图形工作站、单个计算机和服务器等）运行数据挖掘任务。FIU-Miner 综合考虑各种因素（包括算法实现、服务器负载平衡和数据位置）来优化计算资源的利用率。

(4) 有效的程序调度和执行。应用架构上包括用户界面层、任务和系统管理层、逻辑资源层、异构的物理资源层。这种分层架构充分考虑了海量数据的分布式存储、不同数据挖掘算法的集成、多重任务的配置及系统用户的交付功能。一个典型的数据挖掘任务在应用之中需要复杂的主任务配置，整合多种不同类型的挖掘算法。因此，开发和建立这样的计算平台和工具，支持应用领域的数据分析人员进行有效的分析是大数据挖掘中的一个重要任务。

FIU-Miner 系统用在了不同方面，如高端制造业、仓库智能管理、空间数据处理等，TerraFly GeoCloud 是建立在 TerraFly 系统之上的、支持多种在线空间数据分析的一个平台。FIU-Miner 提供了一种类 SQL 语句的空间数据查询与挖掘语言 MapQL。它不但支持类 SQL 语句，更重要的是可根据用户的不同要求，进行空间数据挖掘，渲染和画图查询得到空间数据。通过构建空间数据分析的工作流来优化分析流程，提高分析效率。

制造业是指大规模地把原材料加工成成品的工业生产过程。高端制造业是指制造业中新出现的具有高技术含量、高附加值、强竞争力的产业。典型的高端制造业包括电子半导体生产、精密仪器制造、生物制药等。这些制造领域往往涉及严密的工程设计、复杂的装配生产线、大量的控制加工设备与工艺参数、精确的过程控制和材料的严格规范。产量和品质极大地依赖流程管控和优化决策。因此，制造企业不遗余力地采用各种措施优化生产流程、调优控制参数、提高产品品质和产量，从而提高企业的竞争力。

在空间数据处理方面，TerraFly GeoCloud 对多种在线空间数据进行分析。

对传统数据分析而言，其难点在于 MapQL 语句比较难写，任务之间的关系比较复杂，顺序执行之间空间数据分析效率较低。而 FIU-Miner 可有效解决以上 3 个难点。

大数据的复杂特征对数据挖掘在理论和算法研究方面提出了新的要求和挑战。大数据是现象，核心是挖掘数据中蕴含的潜在信息，并使它们发挥价值。数据挖掘是理论技术和实际应用的完美结合。数据挖掘是理论和实践相结合的一个例子。

作者简介

李涛：博士，教授，南京邮电大学计算机学院院长，2004 年 7 月获美国罗彻斯特大学（University of Rochester）计算机科学博士学位。2004 年至今先后任美国佛罗里达国际大学（Florida International University, FIU）计算机学院助理教授、副教授（终身教授）、正教授（Full Professor）、研究生主管（Graduate Program Director），FIU 计算与信息学院数据挖掘实验室主任，博士生导师。由于在数据挖掘及应用领域成效显著的研究工作，曾多次获得各种荣誉和奖励，其中包括 2006 年美国国家自然科学基金委颁发的杰出青年教授奖，2010 年 IBM 大规模数据分析创新奖，并于 2009 年获得佛罗里达国际大学最高学术研究奖。此文为南京邮电大学李涛教授生前发表过的稿件，未经本人审核（注：李涛教授，因突发疾病，抢救无效，于美国东部时间 2017 年 12 月 13 日在美国不幸逝世，终年 42 岁。）

大数据时代的人工智能

重庆邮电大学研究生院院长 王国胤

近期，大数据和人工智能已成为社会的焦点，受到社会上各阶层及各领域人士的广泛关注。

一、人工智能的诞生与发展

人工智能经过了 60 年的发展，其中经历过很多坎坷，形成了几个流派。近几年，由于云计算和大数据的发展，人工智能赢到了很好的发展机会：云计算为人工智能提供了强大的计算平台，大数据为人工智能提供了丰富的数据资源。因此，人工智能从过去的游戏和科幻，变为如今的产业与现实。

早在 1950 年，“人工智能之父”图灵教授便提出了一个图灵测试的概念，测试某个系统是否具有人类的智能能力，需看系统能否“骗”过人类。如果人不能分辨其是系统还是人类，便认为该系统具备了人类的智能能力。

近一二十年，人工智能在各个领域取得了很多突破性成果。在一些领域中已超过人类，如游戏、人脸识别、语音识别等方面。

2017 年 1 月，人工智能系统 AlphaGo 横扫人类的一流围棋棋手。这对人类棋手而言，是一件不可思议的事情，一个人类棋手是难以奢望实现 60 盘连胜的。未来的人工智能系统（如机器人），甚至可能不再是冷冰冰的系统，而是有情感、有温度的系统。

人类通常认为人是最伟大的。以机器人为例，会认为人比机器伟大，机器只是人的一部分。对机器人可以做一个简单定义：用机器来实现人的一部分能力，这样的系统便是机器人。能扫地是扫地机器人，能做饭是做饭机器人。相反，是否会出现用人来实现机器系统部分能力的情况呢？这个问题是很现实的。如网上的游戏系统，当一个人进入游戏系统时，他也在游戏网络

中为其他玩家提供服务。在此系统中，作为提供服务的人，与系统中提供服务的软件功能相同。因此，人也是机器系统的一部分。

关于 AlphaGo 围棋系统，AlphaGo 战败了人类棋手李世石。实际上，我们看到的是黄士杰在与李世石下棋。在此过程中，可以将棋手黄士杰看作 AlphaGo 落棋的机械臂，也可以将 AlphaGo 看作黄士杰请的高参。AlphaGo 的胜利其实是 AlphaGo 团队的胜利，是人机联合系统的胜利。

二、机器智能会否超越人类？

人工智能的发展已引发了社会的诸多争议和担忧，担心人工智能在未来会超越甚至毁灭人类。其中有 3 种观点：①超越派——机器智能最终将超越人类；②无线趋近派——机器智能会永远接近人类智能，但不会超越；③中立和已经发生派——机器智能与人类智能充分融合。但就个人来看，认为机器智能只是人类智能的高级技术工具之一。人工智能系统战胜人，实质上是人战胜人，是利用先进科技工具的一个人或一群人，战胜了另外一个或一群没有这些科技工具的人。如果某一天，核武器毁灭人类，那是人类毁灭自己。人工智能也是如此。

通常而言，科技用于民生，可以造福人类；也可以用于战争，毁灭人类。人工智能同样是具有双面性的，是一把双刃剑，可以下围棋，也可以建“天网部队”。因此，这样的技术是需要拥有的，也是需要考虑约束对它的使用的。

三、拥抱人工智能科技创新的新春天

追溯历史，人类已经从农业时代发展到工业时代、信息时代，未来的发展将是智能时代。可以看出，从农业时代到工业时代，主要是由于蒸汽机、电力和内燃机等技术的推动，这些技术对全社会产生了重要的影响。从工业时代到信息时代，主要是由于计算机和信息网络等技术的推动，这些技术对各个行业领域都产生了普遍的影响。如今，人工智能技术可以推动各行业领域实现从信息化到智能化的发展，它是一种可以影响众多行业领域的共性技术，未来社会将从信息时代向智能时代发展。信息时代实现的是信息化，实现了对数据的获取、存储、传输和简单处理，未来需要特别关注大数据时代

的智能化研究问题，即大数据时代的人工智能。

获取数据是需要很大代价的。很多领域虽然建立了数据中心，获取到了数据，但对于数据信息还未充分利用。因此，在未来，需要让大数据实现其自身的价值，需要对大数据进行智能分析处理，这也将是未来技术发展的重点。

四、国家战略方向

2015 年，国务院发布的《关于促进云计算创新发展培育信息产业新业态的意见》文件中，已明确提出了进行大数据挖掘分析等关键技术的突破。在国务院发布的《关于积极推进“互联网+”行动的指导意见》文件中，也对大数据技术提出了很高的期望，其中包括 11 项重点行动：①创新创业；②协同制造；③现代农业；④智慧能源；⑤普惠金融；⑥益民服务；⑦高效物流；⑧电子商务；⑨便捷交通；⑩绿色生态；⑪人工智能。其中，第①项是不分行业领域的，具有共性；第②～⑩项是 9 个典型的行业领域；第⑪项是人工智能，是关系未来整个社会发展的一项共性关键技术。通过高频词的分析可以发现，“智能”的频率甚至超过了“大数据”“互联网+”。而且，智能是未来每个行业领域都需要关注的技术。

人工智能技术的发展是具有普遍意义的，可以推动不同行业领域的发展。人工智能本身也作为一个产业领域，受到世界各国的关注，我国在培育人工智能新兴产业方面，注意到了人工智能的核心技术及其在各行业中的智能产品创新。

现代社会已实现了信息化，我们获取到很多数据。如何利用这些数据，是大家都需要考虑的问题，且国家也大为关注，国务院发布的《促进大数据发展行动纲要》文件中，提出了对众多智能数据分析处理技术的支持。

作为一个国家战略性新兴产业，人工智能本身也是受到高度关注的。国家“十三五”规划中也提出要加快人工智能的支撑体系建设，推动人工智能技术在各领域的应用。

在国家实施的战略性新兴产业发展中，能起到最重要、最广泛作用的还是智能技术。智能的出现频率，远远高于其他相关技术，如云计算、生物技术、大数据、新材料、互联网、新能源、制造等。2017 年，我国将实施“人

工智能 2.0”计划。这将推动整个国家从信息化向智能化发展。

那么，生活在互联网时代，你从互联网中看到了什么？一般人都会说看到了很多数据与信息，如今是信息爆炸的时代。但在今天，我们需进一步看到其中的知识和智能，挖掘处理数据，发现其中的知识，为实现大数据的价值服务。

实现大数据的价值，需要对大数据进行智能分析与决策研究。其中，需实现从数据到信息、知识，再到智能的研究，包括一系列技术问题。围绕这一系列问题，我在重庆邮电大学的研究团队已经开展了大数据智能计算研究，分析、发现大数据中的知识和价值，如对信息安全大数据、虚拟现实大数据、环境空间大数据、医疗健康大数据、政务大数据、流程工业大数据、电信运营大数据等的分析研究。

希望 CIO 们在未来能够更加关注大数据的智能处理研究，推动人工智能在各行业领域的应用发展，推动社会向智能化方向前行，为实现社会迈入智能化而共同努力。

作者简介

王国胤：重庆邮电大学研究生院院长、计算智能重庆市重点实验室主任、大数据智能计算示范型国家国际科技合作基地主任、二级教授、国家“万人计划”科技创新领军人才、百千万人才工程国家级人选、教育部“长江学者”特聘教授、教育部“新世纪优秀人才支持计划”人选、中科院“百人计划”专家、重庆市首席专家工作室领衔专家、重庆市学术技术带头人，国际粗糙集学会（IRSS）理事长、中国人工智能学会（CAAI）副理事长、重庆市计算机学会副理事长、重庆计算机用户协会副理事长、重庆市人工智能学会副理事长。主要从事粗糙集、粒计算、大数据挖掘、机器学习、智能信息处理、认知计算等领域的理论及应用研究，出版专著 15 部，发表论文 200 多篇，论著被他人引用 8000 多次，获全国优秀教师、全国高校优秀中青年骨干教师，国家级高等教育教学成果二等奖、重庆市自然科学一等奖、重庆市教学成果一等奖，第四届“吴文俊人工智能科学技术创新奖”二等奖，带领的教学科研团队获评国家级教学团队、重庆高新创新团队、重庆市十大杰出青年群体。

3D 打印格局的大视野认知

北方工业大学机械与材料工程学院副教授 胡福文

目前，“格局”二字比较流行。个人认为，所谓的“格”是认知的分辨率，“局”是不同的认知维度，认知的分辨率加上认知的维度便构成了认知的格局。所谓大视野认知，是从大尺度、大范围来认知。

一、3D 打印的产业链

所谓产业链，是指产业中不同企业间的供给与需求关系，它们自然形成了一个供求链条。3D 打印的产业链向上游延伸便进入到基础产业环节和技术研发环节，主要包括基础材料、打印原材料的研制供应，3D 打印设备的研制和供应，还包括相关软件、数字化设计的专门机构，这便是 3D 打印的上游。

3D 打印产业链向下游延伸便进入到市场拓展环节，主要是与具体的行业应用相结合，如民用消费领域、工业设计领域、航空航天军工领域、服装设计、汽车工业、传播、医疗、饮食等。下游主要是各个行业的具体应用。

在产业链中不同环节的利润率是不同的，这一点对从事 3D 打印行业的人而言特别重要。打印设备的利润率一般为 30%左右，打印材料的利润率一般为 60%以上，而应用服务的利润率可以达到 40%左右，通过这些数据可以看出，打印材料是 3D 打印产业链的核心环节，具有高技术、高附加值的特点。

目前 3D 打印的耗材主要被美国的 3D Systems、德国的 EOS 等大公司垄断，我们国家涉及 3D 打印材料的企业，尤其是高端基础材料的企业相对而言还比较少。因此，3D 打印耗材已成为制约我国 3D 打印产业发展的核心瓶颈之一。

二、3D 打印的市场成长

2015 年统计的 2014 年全球 3D 打印市场规模是 41 亿美元，当时的 2014 年也被称为 3D 打印元年，年增长率超过了 35%。2014 年的统计值是 2015 年发布的，2015 年预测 2016 年的市场规模会达到 70 亿美元。根据 Wohlers 公司 2017 年 4 月发布的最新报告，2016 年全球的 3D 打印市场规模达到了 60.63 亿美元，比预期的 70 亿美元略低，如图 1 所示。

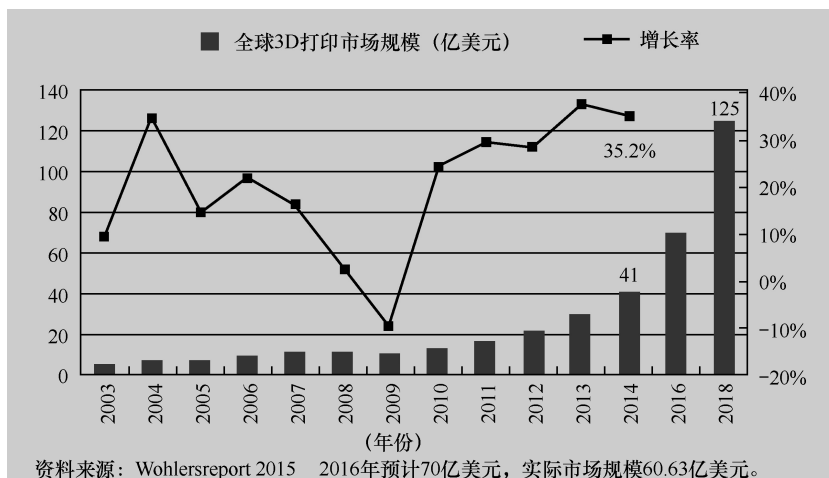


图 1 Wohlers 公司 2015 年发布的全球 3D 打印市场规模报告

3D 打印巨头 3D Systems 2016 年总收入 6.33 亿美元，比 2015 年的 6.41 亿美元略有下降。我国规模最大的 3D 打印企业先临三维 2016 年达到 3 亿元，利润最多 3300 万元。根据我国海尔电器发布的 2016 年业绩，收入约为 638.55 亿元。通过对比发现，3D 打印仍然是一个刚出生的“婴儿”。

如图 2 所示，美国的 Stratasys 占 41.1%，3D Systems 占 15.3%，德国的 Envisiontec 占 10%，爱尔兰的 Mcor 占 5.6%，德国的 EOS 占 2.9%，其余公司占 25.1%。美国的 Stratasys 和 3D Systems 占据了全球将近一半的市场份额，这是有历史原因的。其中 3D Systems 公司成立于 1986 年，它是由发明家 Charles Hull 创立的，是世界上第一家 3D 打印公司，其核心技术是 SLA 光固

化成型技术。Stratasys 是 1988 年由 ScottCrump 创立的，其核心技术是 FDM 熔融沉积成型技术。

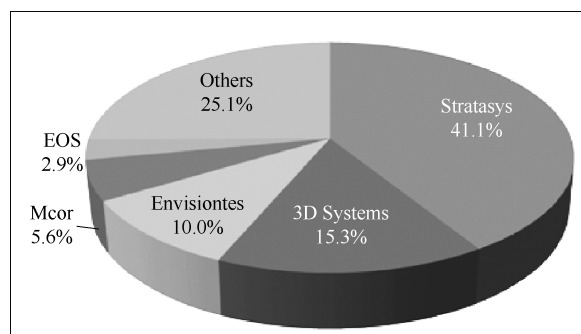


图 2 2015 年全球主要的 3D 打印公司的市场份额分布情况

从 1988 年一直统计到 2015 年。可以看出，世界上 3D 打印机的保有量仍主要集中于美国，占了将近 40%，中国大陆占 9.5%，是发展中国家最高的，也基本上反映了我国 3D 打印在全球的位置。

图 3 所示为截至 2015 年的统计数据。可以看出，光敏树脂占比最高，占了 45%；树脂粉末占了 24.9%；消费级用到的纤维占了 15.1%；金属粉末占了 11%；其他材料占了 3.1%。耗材的分布也反映了不同 3D 打印技术原理的市场分类情况。常见的 FDM 技术耗材占到了 15%。金属粉末近几年在快速上涨，如图 3 所示。

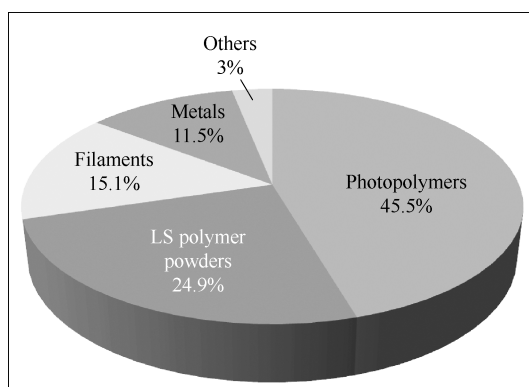


图 3 3D 打印耗材的份额

如图 4 所示，3D 打印的 32.5%用于功能性零件，具有一定的功能，其占比最高。用于配合设计的占 16.2%，用于复杂的装配验证；可视化辅助占了 8.5%，主要是设计师、工程师用于和客户间进行交流，包括医学上的辅助化治疗；教育研究占了 10.1%，包括创客教育和一些科研实验。

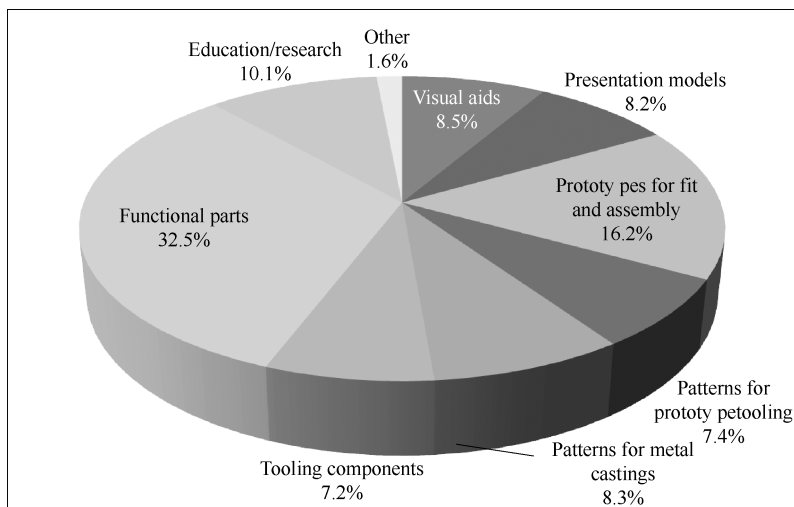


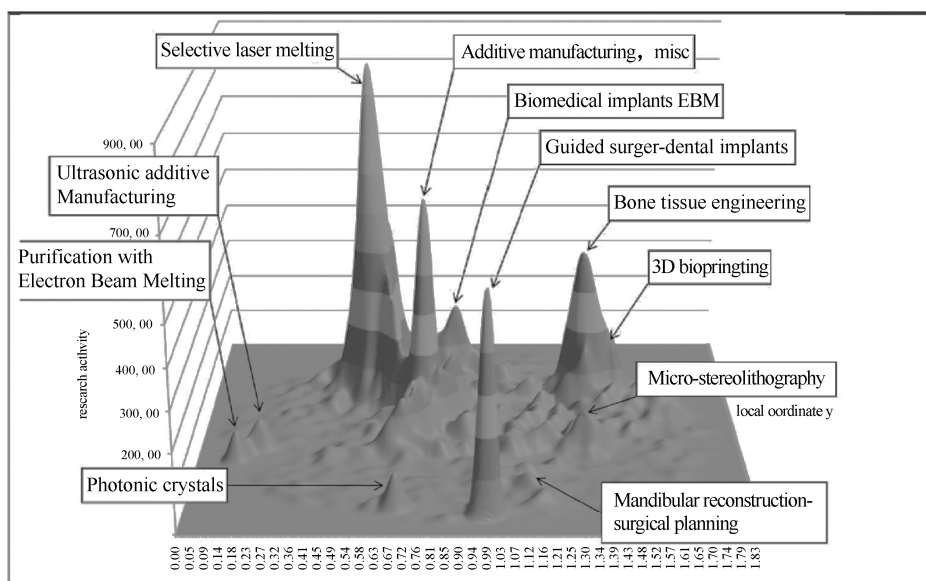
图 4 3D 打印的应用情况（2015 年）

三、3D 打印的技术发展

由图 5 可以直观看出 SLM（选择性激光烧结技术）是当前研究的热点，是和传统制造进行竞争的最后堡垒。

3D 打印在各领域无边界的创新应用、颠覆性应用、奇特应用、混合应用也是当前的研究热点，这说明 3D 打印技术正在向各领域快速渗透，但还在研究层面，这是 3D 打印未来发展的一个成长性空间。

另外，从图 5 中可以看出，3D 打印在一些医疗领域应用全面开花，包括 3D 生物打印、人体植入、组织工程、手术规划导板等都是当前的研究热点，可以说，3D 打印在医疗医学领域有一定的引领特征。



资料来源：Web of Science, own calculation.

图5 根据文献计量分析得到的图

四、总结

通过上述分析，可以得到以下3点基本认知：

（1）3D打印产业与家电、汽车等其他成熟产业相比，仍然是一个刚刚出生的“婴儿”，市场规模还非常小，属于朝阳型的上升期产业。

（2）不论是3D打印技术，还是产业，美国仍居于绝对的领头羊地位，具有一定的垄断地位。通过3D打印设备的保有量、3D打印市场份额的占有量、技术的发展史可以看出，美国在整个3D打印行业具有垄断性地位，而中国的3D打印企业与美国、德国、爱尔兰、日本相比，还未表现出明显的优势。

（3）3D打印技术与行业应用在同时快速发展，正在向各行业快速渗透，包括教育研究行业，教育研究行业可为我们提供一大批具有3D打印思维的人才，他们的成长可将3D打印技术的优势发挥得淋漓尽致。

作者简介

胡福文：北京航空航天大学航空宇航制造工程专业工学博士毕业。现为北方工业大学副教授、硕士研究生导师，中国人工智能学会智能交互专委会委员、中国增材制造产业联盟理事、中国工程机器人大赛委员会委员，主要从事3D打印创新技术、智能机器人技术等领域的研究工作和高等工程创新教育工作。公开发表研究论文30余篇，SCI/EI检索12篇，拥有国家专利和软件著作权8项、编著2部。近4年主持各类科研及教研项目20多项，作为指导教师荣获国家级、省部级科技竞赛奖励50多项。注重教育教学方法创新，提出了“教、学、做、用、研、创”六位一体的创新教育教学模式，荣获北方工业大学高等教育教学成果一等奖。2016年6月被北京市委授予“北京市优秀共产党员”称号。

人工智能——奇点还是支点

北京理工大学自动化学院教授 马宏宾

从 2016 年 AlphaGo 战胜李世石后，人工智能在 2017 年尤为火爆。从 AlphaGo 谈起，前段时间 AlphaGo 与柯洁对弈，可以看到在短短的一年时间内，技术有了突飞猛进的发展。在 AlphaGo 和李世石对弈时还需要 1920 个 CPU 和 280 个 GPU，甚至还需要专门的发电机（每下一局成本是 3000 美元，其中的费用主要是电费），那时的人工智能机器还需要学习上千万张棋谱。2017 年 AlphaGo 与柯洁对弈时变为了单机，用的也不是 GPU，而是谷歌开发的 TPU 技术。到了最近出现的 Alpha Go Zero，已经能够无师自通，不用再读任何棋谱。由此看来，人工智能技术毫无疑问在近年来得到了突飞猛进的发展，其进步速度非常可怕。

目前处于一个非常伟大的时代，面临着技术的诸多变革。在人类历史上，技术的发展很大程度上是由一些勤奋的懒人推动的，因为人类总是向往美好的生活，希望能过上舒服的日子。由于这些想法，人类能从远古的时代学会造工具、取火，逐渐将人类和动物区分开。在人类的发展历史上，之所以能步入到当前的智能时代、大数据时代，很大程度上是因为人类对未来抱有很多美好的想法，逐渐发明各种技术、机器、机器人，将人类的各种功能解放出来，进而推动未来的发展。

一、人工智能的现状

近年来，人工智能的发展很大程度上与深度学习取得的突飞猛进的发展有关，深度学习在传统的 CPU 上进行计算时，需要海量的数据集，训练所需要的计算量非常大，传统的 CPU 可能需要非常多的时间。GPU 作为传统的显卡，本身具有并行计算能力。新时代下，人们发现 GPU 不仅可以进行单纯的

图像渲染、计算，还可将其作为计算模型，将人工智能的算法在 GPU 上实现，速度会变得很快。因此，GPU 在近年来得到了很大的关注。英伟达公司的 GPU 是全球知名的，2016 年上市公司的股价上涨了 5 倍。这反映出背后人工智能需求的巨大变化。

二、人工智能的历史背景

很多人认为，1956 年的达特茅斯会议产生了人工智能，于是将这一年称为“人工智能的元年”，但不论从人工智能的研究，还是人工智能本身而言，1956 年都不是一个起点。实际上，“人工智能”（Artificial Intelligence）这个词早在 1955 年 8 月麦卡锡、明斯基、罗切斯特、香农等人提议举行达特茅斯会议时就已经提出来了。普遍的误解是“人工智能”这个词是麦卡锡想出来的，但麦卡锡晚年回忆说，一直有印象“人工智能”这个词最早是从别人那里听来的，由于当事人都已经仙逝，这个词的来源已难考证。

换一个角度，站在人类发展的历史长河中来看，人工智能的思想或梦想来源已久。从几十万年前人类诞生以来，在演化过程中人类自身有各种各样的局限性，人不希望自身去做一些工作，包括人在与大自然搏斗的过程中，人的能力和野兽 PK 时有着很大的缺陷。带着这些问题，人不断地产生智慧，通过集体的协作，想办法发明工具，通过外在的工具将人类的能力解放出来。因此，可以认为，人类的发展史某种程度上便是人工智能的发展史。人类历史发展的进程越来越快，呈显著加速发展状态。目前我们正处于技术革命爆发的一个关键阶段。

人工智能其实有诸多先驱，阿兰·图灵、冯·诺依曼、香农、维纳、赛弗里奇、丘奇等都功不可没。我们重点说一下图灵，他的人生经历充满传奇，令人唏嘘不已，他提出了图灵机的计算模型，成了后来计算机技术的基础；他发明的破密码机器在“二战”中为以英国为首的盟军取得胜利立下了汗马功劳。计算机界的最高奖——图灵奖以他的名字来命名，他在 20 世纪 40~50 年代提出图灵测试的基本想法，如果机器和人类能展开对话，且人类无法辨别其机器身份，这就是所谓的机器智能。目前来看，应该还未能达到这样的标准。

毫无疑问，人工智能已成为一个备受关注的重要领域，潘云鹤院士等专家也提出了人工智能 2.0 的概念，并推动国家出台了新一代人工智能战略规划。“人工智能 2.0”这个概念是被热炒的，但从现实看来，大部分已上市的机器人的智能水平与人们的期望还有一定的差距。

人工智能在历史上从未像如今这么火爆过，每次新技术的出现，人们都对其抱有很大的期望。事实上在人们逐渐应用新技术的过程中，人们发现很多技术并不能解决实际的问题，如疾病诊断。当时的人工智能技术并未达到很高的程度，从波峰重重跌落下来，成为被人们唾弃的对象，这样的历史是值得反复回味的。

在人工智能的历史上，人类的历史与幻想、科学发展、猜测和想象有着密不可分的关系。除了达特茅斯会议后的发展，其实在那之前，人们很早便想制造自动人偶，来模拟人类或者帮助人们劳作。我国历史上就记载了许多机器人偶的故事，如《列子》中就记载了西周穆王时工匠偃师造人的故事。在西方，相传 1773 年，雅克德罗发明了 3 个人偶，分别是画家、作家、音乐家，但那时是纯粹机械的，这些人偶能做一些人能做的演示，可见人工智能模拟人的探索由来已久。

（一）在人工智能历史上起到重要推动作用的 3 个“C”

1. 计算（Computation）

1946 年，美国的数学家莫克利和艾克特制造了世界上第一台通用计算机 ENIAC，电子数字计算机的发明是人类历史上一件非常重要的事情。电子数字计算机俗称电脑。电脑的发明，为人类带来了天翻地覆的变化。

1950 年，图灵在论文中提出了机器思维的问题，他是一个天才的哲学家、数学家、思想家，他考虑机器能否模拟人类，像人一样思考，那个年代脑科学还不够深入，如何定义思考是很困难的问题。在此背景下，图灵引入了“模拟游戏”的概念，具体是指看一个系统或一个机器是否是智能，就让人与其对话，中间是隔开的，人与机器在对话过程中如果丝毫察觉不到彼此间的差异，便通过了图灵测试，具备了足够的智能。这也是人工智能梦寐以求的一个目标。计算机科学领域国际最高奖以图灵命名。图灵提出了图灵机的构想，计算的模型通过一些简单的自动机迭代，能完成任意的计算，将抽象的问题

转换成通过计算来表达任意的可能性，图灵计算模型成为计算机领域非常根本的一个模型。

除了著名的图灵机模型，实际上，在现实的计算机程序设计中，图灵的导师——数学家丘奇的理论在起着非常大的作用，丘奇的理论让很多事情变得简单，而图灵的机器却过度复杂。丘奇提出了 **Lamda** 演算，这是几乎所有程序设计语言的理论基础。丘奇后半生致力于用数理逻辑把常识形式化，影响了很多，他的学生克门尼是 **BASIC** 语言之父，在达特茅斯会议前两年担任达特茅斯学院数学系主任；丘奇从母校普林斯顿大学挖来了麦卡锡等人，而麦卡锡是达特茅斯会议的主要召集人。图灵奖得主中有大概 1/3 的人从事了与程序设计语言相关的工作。因此说，在计算技术的发展中，大批数学家起到了重要的作用。

2. 通信（Communication）

数字通信的奠基人是香农，他于 1948 年发表了具有深远影响的论文《通信的数学原理》，从而被人尊为信息论之父。他是一位数学家，首次以数学的方式描述信息的传输，给出了信息熵和比特的概念。如今，赛百空间传输、沟通交流、分享的内容都是数字化的，都是通过比特信道进行传输。香农在互联网出现之前便建立了信息通过信道如何传递、信号在传递过程中会有噪声、会受到污染、由于信道的不稳定性导致传输信号有可能不准确、如何实现可靠的通信等疑问，于是他提出了信息论的三大基本定律。目前，通信技术发展至今，都在逼近香农当时描述的极限，并未突破信息论的框架。香农开启了数字通信时代，通信使得世界可以连接在一起，后来的互联网让世界上的计算机可以连接在一起，移动互联网让世界上的手机等移动终端可以连接在一起，物联网则有可能实现万物互联，这些伟大技术革命的源头都可追溯到香农的贡献。香农也直接参与了达特茅斯会议的筹备。

3. 控制论（Cybernetics）

控制论是人类历史上另一个非常伟大的科学进展，物理、化学、生物等自然科学首先是用来认识世界的，而控制论则是人们改造世界的横断性思考，控制论的模型可以用于生物、机械、化学、电子等领域。因此，同一个数学模型在各个领域都有可能得到应用。经典的“控制论之父”是维纳，他是一

位数学家，他很希望为国家做贡献，第一次世界大战时积极要求参军，但被国家拒绝了。后来，随着维纳的研究，接了一些军方的项目，于是产生了很多控制论的想法，比如维纳滤波就源自弹道轨迹的估计。不过值得一提的是，维纳晚年拒绝接受军方的任何资助。维纳 1948 年出版的《控制论》一书中有一个副标题是“人有人的用处”。在维纳提出控制论时，他已经意识到控制论是在人、动物所有的世界中的横断性内容，人对其研究有了反馈的思想，便可以做出各种各样的决策与行为，进而对世界产生影响，那时维纳便意识到生物界、动物界，包括人造的系统中，反馈控制是一个基本的概念，“人有人的用处”一开始便强调了技术应该为人服务，人应该放在更大的系统中，作为其中的一个环节。维纳是控制论之父，他的理论产生了革命性的影响，比如，维纳《控制论》一书的第一个读者赛弗里奇是模式识别的奠基人，他写了第一个可工作的人工智能程序。历史上苏联在 20 世纪上半叶关于人工智能的研究基本都在控制论的名义上进行。现代控制论的出现也归功于一个数学家——卡尔曼，卡尔曼于 1960 年左右发表的几篇论文产生了深远的影响，阿波罗登月计划就极大地得益于卡尔曼滤波延伸出来的扩展卡尔曼滤波。

（二）在人工智能历史上起到关键作用的人物

在人工智能历史上起到非常大作用的一些开创性人物，从上面的介绍中可以发现所有的这些人物都是思想家、哲学家、数学家，有很好的数学功底。有人说数学家是不食人间烟火的，当他们将思维升向宇宙最根本的问题上时，甚至结合一些具体的问题时，便能产生巨大的威力。

除了前面重点介绍的图灵、香农、维纳，这里要特别强调一下伟大的数学家冯·诺依曼。他其实一开始只研究纯数学的一个分支——代数结构，但由于第二次世界大战时国家的需要，他才开始将目光转向一些与实际背景有关的领域，做出了开拓性的伟大贡献。比如，虽然 ENIAC 不是冯·诺依曼发明的，但冯·诺依曼提出了二进制，用 0 和 1 表示所有的信息，ENIAC 当时并不是基于二进制的，它需要很多按钮来控制，占地面积非常庞大，相当于几十间的办公室，每秒都有比如真空管等零件产生发热问题出现故障的可能性，其计算性能还远远达不到目前的计算机性能。冯·诺依曼看到 ENIAC 后提出计算机可以引入一种存储程序，不需要每步都由人来操作，完全可以将

人的工作编码后变成程序，放到计算机内存中，目前用到的计算器还未脱离冯·诺依曼提出的架构。冯·诺依曼这一构想在计算机历史上起到了非常重要的作用。冯·诺依曼还是现代计算机体系架构之父，他提出的存储程序的思想，至今几乎所有的电子计算机都在采用；此外，多数人不知道的是，冯·诺依曼还是神经网络的前驱，他启发了明斯基等人的工作，现在大热的人工智能主要是由于深度神经网络的成功应用，因此，我们应该纪念冯·诺依曼等人的开拓性工作。

冯·诺依曼也是博弈论、运筹学之父。博弈论也叫 Game Theory，很多诺贝尔经济学奖获得者都是研究博弈论的，博弈论的创始人就是冯·诺依曼。冯·诺依曼与其合作者在 1949 年写了《博弈论与经济行为》，这一门新的学科把人和人之间、军队和军队之间的所有涉及现实的多方决策都引入矩阵对策的数学模型中了。AlphaGo 与围棋高手的对弈属于顶级高手间的博弈。博弈论也是人工智能技术的一个重要支撑。

对人工智能历史起到关键作用的大多是数学家。例如，中国第一届国家科技奖获得者是吴文俊先生，目前人工智能领域国内设的最高奖项是吴文俊人工智能奖，该奖是以吴文俊先生的名字命名的，吴先生对人工智能也有非常大的贡献，他在 20 世纪 40~50 年代对博弈论的一些根本问题进行了研究。因此，当我们回顾历史时，人工智能的历史本质上与数学、计算机、逻辑及各个领域的应用是密不可分的。

（三）达特茅斯会议报告方向

所谓的人工智能元年是 1956 年，达特茅斯会议的发起人麦卡锡、罗彻斯特、香农当时在 1955 年写了一个报告，报告的方向有以下几个方面：

（1）**自动计算机**。研究计算机用自动编程的方式实现自动的计算。

（2）**编程语言**。为了使计算机能为人工作。

（3）**神经网络**。比如深度学习的背后便是神经网络，早在 1955 年的达特茅斯会议上便被列为一个重要的研究重点。

（4）**计算规模理论**。即计算复杂性，算法完成某项任务需要多大的计算、如何优化计算、计算资源应满足什么要求等问题。

（5）**Self Improvment**。也就是机器自身学习能力，机器能否利用一些新

的数据或信息更新自己的系统，深度学习属于机器学习的根本性思想。

（6）**抽象**。这是人类的一种能力，人类之所以能思考，就是因为人类具有抽象的能力。人对宇宙万物进行梳理，总结出一些规律与规则。抽象是一个非常哲学、非常心理学、与人类思维意识非常有关联的词。在 1955 年达特茅斯会议的筹办者心目中，人类的思维、人类的意识、人类的逻辑抽象等能力早已被纳入到议程上。

（7）**随机性和创建性**。那时的先哲便意识到为了真正研究所谓的智能，赋予机器人类的智能，不可避免要考虑到随机性因素，很多事情是不确定的，如人的突发奇想某种程度上是随机的。与随机性相关联的是创造性，创造性与随机性有一定的关系，生物学上如基因突变，正因如此才可能保证生物的多样性。

因此，从上述七大关注领域来看，半个世纪前的先知们已经意识到，对于如何理解人类的智能有一些关键因素，当时都已纳入考虑范围。到 1956 年达特茅斯会议召开时，据历史考证当时的参会者不止 10 人，摩尔、麦卡锡、明斯基、赛弗里奇、所罗门诺夫、香农等科学家都有不同的背景，有的是数学家，有的是科学家，有的对心理学感兴趣，有的对逻辑感兴趣，有的对人类大脑感兴趣。这些人后来都成为人工智能领域或计算机领域卓有成就的大家，明斯基、麦卡锡、纽厄尔、司马贺这几个人后来都获得了图灵奖。

（四）人工智能的相关预言

回顾这段历史，天才的哲学家、数学家、思想家其实很早便意识到，如果能发明一些像人一样聪明、智慧、能干的机器，将会为人类社会带来巨大的改变。因此，很多人投入了毕生的精力和努力。在此过程中，随着技术的发展，很多人做出了各种各样的预言。

1957 年，达特茅斯会议的参加者司马贺预言，十年内计算机下棋会把人击败，当然，这个预言在十年后未能实现。

1968 年，麦卡锡和象棋大师列维打赌称，十年内下棋程序会战胜列维。当然，麦卡锡输掉了这个打赌。我们知道，计算机或人工智能系统战胜象棋大师一直到 1997 年，从 1957 年司马贺的语言开始到 1997 年大概过了 40 年，计算机的人工智能系统才赶上人类象棋大师的水平。因此，麦卡锡为此输掉

了 2000 美元的赌注。

1968 年，达特茅斯会议的参与者明斯基预言，30 年内机器智能可以和人类一决高下。当然，这个预言到期的 1998 年，人工智能还未像现在一样引起广泛关注，机器智能的水平和解决的问题有一定的局限性。

（五）科学家对人工智能的看法

人工智能这样的一个广为街谈巷议的概念是否有一个普遍认同的定义？很多说法各不相同，到目前为止，还存在很大的争议。大家普遍都在谈，但实际上有多少人懂？全世界伟大的科学家都表现出不同的看法。例如，霍金便认为人工智能会给人类带来非常大的灾难，甚至表示担忧。另外，真正在一线距离人工智能较近的技术工作者并未表现出很大的担忧。可以说，智能时代的人工智能技术已取得很大的进展，未来还有很长的路要走。目前来看，发展较好的是弱人工智能，这样造出来的机器并不是说能像人一样思考，有创造性，只是看起来是智能的，但不是真正拥有智能。例如，物体识别、人脸识别、语音识别等都属于弱人工智能的范畴，这些领域已取得了比较大的发展。

在此背景下产生了各种各样的讨论，比尔·盖茨提到，低智力的人工智能在不远的将来将变成积极的劳动力替代工具，他也担心超级智能系统未来会强大到足以变成一个忧患，他补充道无法理解为何有人对此漠不关心。

霍金作为一个非常伟大的天才物理学家，他相信人工智能极可能是奇迹，又可能是灾难，他将人工智能称为人类历史上最大的事件，这项技术如果利用得当，将有助于人类消除战争、疾病和贫穷，但人工智能的爆发性增长也可能取代人类的金融市场，变得比人类研究人员更有创造性，摆脱人类领导者的控制，甚至开发人们不能理解的武器，霍金称人工智能很可能是人类历史上最后的事件，除非我们学会如何规避风险，他对人工智能的发展表示出非常大的担心。

有人认为获得更强的计算能力只是时间问题，随着摩尔定律的继续推进，计算资源将越来越便宜，人工智能自然会超过人类，扎克伯格认为这种说法是不正确的，事实上并没有从本质上理解通用的学习原理，他表示我们不应害怕人工智能，相反，我们应该期待它会给世界带来许许多多的好处。

吴恩达之前是百度深度学习研究院的首席科学家，他曾这么说：“其实有一句话我在一年前讲过，其实对超级人工智能的担心，就像担心火星人口过多，可能几百年后有人在火星，那时的火星人很多，环境受污染，疾病蔓延，那时可能会真的担心，但目前是未知的，其实很多人还是不太理解人工智能什么可以做，什么不可以做。”

三、人工智能的未来

（一）对人工智能的担忧

目前，类脑智能的研究比较火热，包括未来如何真正将人脑的机理运用到发展人工智能。未来强人工智能或一般人工智能的发展方向也是人们争议的一个领域。当然，也有一些科学家对超人工智能表示担忧，甚至有新闻说人工智能将在 2029 年超越人类。

实际上，人工智能是达不到人的一些能力的，如创造力、直觉、审美、悟性、洞察力等。个人认为，目前的人工智能技术本质上是一种优化，在一个非常大的数字空间中，通过一些想法或技术实现优化。这种监督式的机器学习其实非常类似于原始的函数拟合问题，维数（数据）更多的话是线性的分类问题。后来的支持向量机、统计学习方法、神经网络、深度神经网络等，都是某种数学的模型，最后对标记过的数据在非常大的参数空间中寻找一些参数，构建数学模型，对数据、物体进行分类或表达。还有一些其他的技术，如决策树、贝叶斯网络、过去的专家系统等都是人工智能的传统方法，这些方法曾引起极大的关注，人们对其抱有很大的期望。例如，作为一个智能医生帮助人们诊断疾病，其初衷是好的，当面临实践中的不确定性、复杂性时，很多内容无法用逻辑量化，很多内容是模糊的，最后人们发现这是达不到期望的。

目前的人工智能热是否会出现实际与期望相悖的现象也是不可排除的，未来人工智能有可能会逐渐沉淀下来。当人们不再关注时，人工智能技术会不断发展，找到更多的应用场景，不断完善。如果人工智能像电和空气一样无形之中融入到人们的生活中，那便可以达到很好的发展水平了。

人们对于人工智能的担忧表现在自己的工作在未来是否会被替代。例如，

操作工越来越多地被机器人替代；无人驾驶商业化后，传统的司机也在逐渐被取代；客服正在逐渐被智能的语音客服所替代；翻译领域也出现了翻译机。可以看出，人工智能将来真的会改变很多人的命运。

对教师而言，传道授业解惑是有一定技术含量的。将来是否会被人工智能技术、大数据、网络改变？这种趋势已逐渐呈现，如公开课、在线教育平台等，可以自适应地根据每个人的学习情况、学习进度，制定不同的学习方案，这便应用了人工智能技术。甚至有一些学者将人脸识别、表情识别与教育的进程相结合。

很多行业都将面临新的挑战，包括新闻记者、客服、金融投顾、行政人员等，都有越来越多的机器及人工智能系统帮助人来做。含金量比较高的编程工作可能需要一定的创造性，随着快速开发工具、平台、自动化编程工具的出现，相对简单的工作有可能会越来越多地被技术替代。

人工智能未来可能会在很多行业都有应用，甚至会拓展出一些新的应用。按李开复所说，50%的行业都有可能被人工智能改造，这说明人工智能未来有可能会产生一场革命性的影响。一场新的技术革命到来后，很多人的工作有可能会受影响。

（二）奇点临近

2005年，预言家库兹韦尔写了《奇点临近》一书，书中指出2045年会成为人工智能的奇点。数学上的奇点概念是指奇异的点，在不同的场合有不同的定义。对于人工智能的奇点，资料中是指人工智能全面超越人类的时间点，这个定义有点吸引眼球的意味，从学术角度来看，具体如何定义“人工智能超越人类”是很难的。

一直到2015年，经过了半个多世纪的发展，人类仍然没有一个明确的答案。因为人类的思维过程从来没有任何一个机器能够完整、全面地模拟出来。像鸟会飞一样，历史上人类会为自己安上翅膀，从高山上跳下来，为科学做了贡献，成为先烈，甚至很多人的名字在科学史上并未记载。飞机的方式和鸟的翅膀并不相同，这是人类的智慧所在，最终能将希望实现的功能模拟出来，但模拟的原理并不完全仿照生物学的原理，仿生学借鉴生物的想法开发新的算法，解决一些问题。例如，蚁群算法等就是借鉴生物学的思想在优化、

搜索人工智能中的想法。因此，这些最终被人类发挥创造赋予机器编程程序和数学模型，机器展现出一定的智能，归根结底是人类赋予的智慧。

2017 年软银孙正义预言，2018 年将是人工智能爆发的奇点。从 AlphaGo 看出，2016 年人工智能已经在中国爆发。同时，赛迪总裁孙会峰预言，2030 年将是人工智能的奇点。

人们对人工智能的想象力是非常丰富的，一旦展现出威力便会对其作用抱有很大的期待，甚至是无限制的夸大，会存在弱人工智能、强人工智能、超人工智能的争辩，这种问题是仁者见仁、智者见智，从人类的终极目标来看，人类希望能造出真正解决问题的、有自我意识、能像人一样思考的机器，这种想象成为很多科幻电影与小说的主题。但另一方面，针锋相对的观点认为，人类不可能造出真正像人一样推理解决问题的机器，机器只是看起来智能，并不具备真正的智能，也没有真正的自主意识。个人认为，人类将机器制造出来，机器实现的智能只是看起来是智能的，要完全实现人类的思维和智慧，实际上是很困难的。公众眼中认为人工智能已经很厉害了，实际上人工智能在实际应用的许多场景中离人们的想象还相去甚远。假如我们真的认为人工智能有朝一日会超越人类，那么，如果人造的机器产生出一些不好的想法，像许多科幻片里那样来试图控制人类，在此背景下的人工智能，有可能不是改善人们的生活，而是一场灾难。

对人工智能而言，我们应该是什么看法？这一奇点是否会真的到来？是否会替换人类的工作？是否会使人陷入很悲惨的境地？这些问题是值得我们认真思考的。

（三）人工智能有没有创造力？

以 AlphaGo 为例，它是一个程序算法，只能用来下围棋，本质是一个搜索空间，计算机的计算能力很强了，但对于围棋的搜索空间可能性而言能力还远远不够。“深蓝”则借助其强大的计算能力，甚至采用了暴力搜索的手段，对各种可能性进行评估。当然，只要将计算机造得足够强大，最终战胜人类是不用怀疑的。“深蓝”是国际象棋电脑，搜索空间比围棋小得多，这也是 AlphaGo 相继战胜李世石、连胜 60 场、战胜柯洁，给人带来非常大的震撼的原因。围棋的搜索空间非常大，这也是它的困难所在，即便如此，AlphaGo

算法的本质还是通过一些技术手段和天才想法缩小搜索空间，做一些优化，优化的背后还是计算。2017 年的 AlphaGo 程序并不是由 2016 年的 AlphaGo 演化出来的，其中主要依赖背后团队的智慧。2016 年的 AlphaGo 下棋需要很多棋谱，计算资源依赖很多台机器、GPU、发电机实现庞大的电力供给。2017 年的 AlphaGo 采用了新的架构及谷歌开发的新的计算单元，使得效率大大提升，一个单机加上两个 GPU 便可以进行预算，无须从海量棋谱中反复学习，人类为其引入了新的想法，AlphaGo 围棋本身是没有演化能力的，因为它不具备真正像人一样的思维。AlphaGo 团队中有会专业下围棋的，但远远达不到 AlphaGo 战胜李世石、柯洁的水平，团队开发的程序能战胜，原因在于人类的智慧。从功耗等方面来看，人类的大脑和 AlphaGo 相比有很多优势，人类智慧的积累、灵感都是特有的。以人类发明汽车为例，一开始人类是很恐慌的，随着汽车的逐渐应用、增多，人们也开始接受这一现实，马车夫的工作则逐渐消失，汽车司机则成为新的职业。未来，随着自动驾驶和无人驾驶技术的发展与普及，社会也会发生新的变化。

因此，人类并不需要对人工智能奇点感到害怕。人类有些东西是人工智能所没有的，如直觉、灵感，很多时候并不是从学到知识后必然性推出的结果。其实任何一个国家的教育制度及教材都是有完整体系的，授课内容、教师也是相同的，为什么会产生差异？因为创造，没有创造，就不可能有丰富多彩的世界。人类的灵感、想象力、创造力等显然不是训练出来的。例如，著名化学家凯库勒睡梦中梦到蛇咬住自己的尾巴，受到启发发现了苯环结构；阿基米德在浴池中发现了浮力定律，这些都是在偶然的机下由于人的灵感产生的。

四、人工智能的典型特征

（一）人工智能系统的内在本质

（1）**由人类设计**：人工智能系统是由人类设计的，其运行或工作必然是按照人类设定的程序或逻辑，因此，人工智能系统从根本上应该离不开人类的掌控。

（2）**为人类服务**：人类创造出来的人工智能系统，在理想情况下必须体

现服务人类的特点，而不应该伤害人类，特别是不应该有目的地做出伤害人类的行为，这是关于人工智能系统应用的基本伦理要求。

（3）体现为机器：人工智能系统本质上是运行了人类设计的程序的机器，按照人类编排的程序模式而工作，因此，从根本上说仍是无意识的机械的物理的过程（哪怕它被设计为模拟了人的某种情绪或功能），其内在本质仍然是被算法、技术、数据等武装起来的机器。

（4）本质为计算：人工智能被称为智能程序的科学，任何一个人工智能系统都借助具有计算功能的硬件运行了一些人为设计、编排的程序（软件），本质上通过逻辑计算或复杂的数学运算，来实现对人类期望的一些“智能行为”的模拟。

（5）核心为数据：人工智能系统的程序会涉及许多信息的计算与流动，在此基础上实现对知识的采集、表达和学习，而信息与知识本质上都是通过数据的处理、分析、计算、可视化等来实现的，因此，数据成为现代人工智能系统的核心，谁掌握了数据的入口，谁就更有可能在人工智能的应用中抢占先机。

（二）人工智能系统的外在联系（“五能”）

（1）能感知环境：随着物联网、智能终端和传感器等技术的不断发展，智能终端或节点（包括各种各样的传感器）未来有可能无处不在，再借助云计算或端计算，可使得人工智能系统普遍具有一定的感知环境的能力，这为人工智能系统在各个领域的应用奠定了基础。

（2）能产生反应：在感知环境的基础上，人工智能系统可对外界输入产生必要的反应（可表现为文字、图像、语音、动作等），进而影响环境，这种反应能力也是人工智能系统与外界联系的一个基本形式，是人工智能系统能够服务于人类的一个关键要素。反应的类型可以多种多样。

（3）能模拟人类：人工智能系统作为机器，首要功能是能够在一定程度上（包括但不限于从结构上）模拟人类的功能，如“看”“听”“说”“走”“做”乃至“想”。

（4）能与人交互：人工智能系统作为被人类发明设计出的机器，必须具有一定的人机交互能力（按钮、键盘、鼠标、触屏、语音、手势、体态、表

情、力反馈、虚拟现实等），这样人工系统才会体现较好的用户体验。

（5）能与人互补：人工智能系统不局限于仅仅模拟人类，还应该与人类产生优势互补，帮助人类去做人类不擅长、不喜欢但机器能够完成的工作（简单智能），而人类则适合去做更需要创造性、洞察力、想象力、灵活性、多变性乃至用心领悟或需要感情的一些工作（复杂智能），而不仅仅是人能做什么，机器就能做什么，这是比较困难的。

人工智能系统计算机的本质是计算，所有的人工智能系统的价值不是抢走人类的工作，人工智能的发展体现在某方面代替人的工作，将人类从重复的、乏味的工作中解放出来，这是人类未来与机器和谐存在的场景，也是人们所期望的。

（三）人工智能系统的关键特征（“五有”）

（1）有适应性：人工智能系统理想情况下应具有一定的自适应特性，即具有一定的随环境或数据或任务变化而自适应调节自身参数的能力，例如，理想情况下的人脸识别系统应能在不同光照条件下都能工作，具有较好的鲁棒性。

（2）有学习能力：人工智能系统理想情况下应具有一定的学习能力或机制，这种学习能力或机制往往是使得系统具有适应性、鲁棒性的一个关键，例如，很多人工智能系统需要对不断更新的数据集进行定期或非定期训练来改善系统性能以应对不断变化的现实环境。

（3）有演化迭代：人工智能系统的发展离不开不断的演化迭代，不可能存在一开始就非常完美的人工智能系统，一个实用的人工智能系统往往是在不断的演化迭代中逐渐完善，到一定阶段后才真正具有对人类的服务能力。人类决定着如何赋予人工智能系统新的算法或机制，机器在人类赋予其学习机制下，可以不断收集新的数据，更新其中的模型，实现功能表现上的演化迭代。

（4）有连接扩展：在智能互联时代，人工智能系统将会具有越来越多的连接扩展，如与云连接、与端连接、与人连接，乃至与可数字化的万物连接，从而未来将更加呼唤开源开放的创新平台，实现依托产业链、生态圈的开放式创新。

（5）有多样应用：人工智能未来有望成为全世界的基础设施，逐渐改变或影响各行各业，其应用的多样性将难以想象。目前，人工智能系统已经开始广泛用于识别图片的元素、实时进行语言翻译、语音控制、智能家居、广告推荐等。

这 5 个方面是使人工智能非常有用的关键特征，不断根据数据、知识库进行更新，能力越来越强，可解决各种问题，演化迭代就像 AlphaGo 一样，背后是人类的智慧，通过人类的智慧产生新的演化迭代，积累新的数据，不断更新，连接到互联网、云上，通过万物互联进而产生应用。

（四）人工智能系统的根本局限（“五无”）

（1）无创造能力：人工智能没有人类的意识所特有的能动的创造能力。由于人工智能系统是人类来设计的，目前的人工智能系统根本上都体现为电脑的能力。人脑与电脑的关系总是人脑的思维在前，电脑的功能在后。所以，人工智能系统的能力完全受限于人脑思维的深度与广度，其输出依赖于其“学习”过的大量样本和人类赋予其的学习机制。

（2）无直觉能力：人类的直觉或者顿悟等能力在人类的意识中起到重要的作用，然而，人类的这些能力显然并非是通过机械的计算来实现的。

（3）无通用平台：人类的智能主要依靠人类的大脑来实现，可以说人类的任何行为和智能，都借助于大脑这一体积不大但功能极为强大的通用平台。人工智能系统则需要针对具体的应用场景和智能要求，来进行相应的算法设计、特定的数据集训练及专门的功能开发，不存在一个通用的平台能同时满足所有应用中关于人工智能的需求。

（4）无人类常识：通过“IF-THEN”规则库、谓词逻辑和专家系统，人工智能系统能够实现一定的逻辑推理能力。然而，万能的逻辑推理体系从根本上来讲是不可能存在的，人类在长达数十万年的演化史和几千年的文明史中积累起来无数的知识与常识，这些知识难以在现有的计算机系统中进行结构化地表达、存储与学习。

（5）无人类情感：人工智能系统可以模仿人的表情和腔调，但这些并不等同于情感交流，也不真正具有人类社会所拥有的社会性。人类社会的情感交流往往建立在信任和真诚的基础上，但人工智能系统的本质仍然是机器，

机器可以帮人干活或者延伸人的能力，但很难使人的内心中产生真正的温暖。

整体来说，人类具有“知识越多，智能越高，反应越快”的特点，但基于状态空间搜索的人工智能系统却受制于“组合爆炸”和维数灾难，呈现“知识越多，训练越长，反应越慢”的特点。人工智能系统不同于人类，可作为人类的延伸与增强，不应该成为人类的完全替代。

五、人工智能是奇点，还是支点？

根据杠杆原理，如果科学原理能恰当应用，致使人类发明一些想法、技术、科学知识，实际上是可以为人类造福的。

目前来看，其实很多对于人工智能的担忧是不必要的，但我们要为人工智能加上一些伦理。我们希望人工智能技术最终目的是为人类带来很多便利，极大提升很多领域的效率，使人类的生活更加美好。如果本着这样的目的，人工智能技术相关领域的发现、研究、投资、应用为人类带来便利，作为一个支点撬动人们的生活、生产及国家的经济，使人们在未来的社会通过人工智能技术对机器发号施令，更好地面对自己，面对心灵的世界，用更少的工作时间实现更好的效果，这是人工智能将来能带来的美好的世界。

人工智能技术带来非常多的应用和挑战，我们更愿看到人工智能得到更好的发展，当然，背后的发展都是由人来推动的。我们需要不忘初心，不要只满足人类的贪婪欲望，不要只想着如何用技术为个人谋福利，而应该利用这些技术推动社会、推动公众乃至整个国家的发展，希望更多的人投入到人工智能的研究应用之中，而不仅仅是泡沫化的癫狂状态，脚踏实地地做一些具体的工作，对其进行细分，与具体的应用结合起来，解决具体的问题，进而为人类带来更多的方便。

作者简介

马宏宾：北京理工大学教授、博士生导师。以“适应·学习·认知”为中心开展研究，探索无人车、机器人及无人机的应用，研究兴趣包括自适应估计与控制、组合导航与智能导航、人机智能交互、机器视觉及机器学习、多智

能体及无线传感器网络应用、嵌入式系统及软件开发、工业大数据。2001 年毕业于郑州大学系统科学与数学系，获学士学位；2006 年毕业于中国科学院数学与系统科学研究院，获理学博士学位；2006 年 3 月至 2006 年 6 月在贝尔实验室工作；2006 年 8 月至 2009 年 7 月在新加坡国立大学淡马锡研究所任研究员；2009 年 8 月至今为北京理工大学教授。入选教育部新世纪优秀人才支持计划，获吴文俊人工智能科学奖、霍英东高等院校青年教师奖、北京理工大学教学成果奖，获北京市优秀人才培养资助计划支持。发表 SCI/EI 检索论文近百篇，与合作者出版英文专著 *Advanced Technologies in Modern Robotic Applications* 与中文专著《矩阵代数、控制与博弈》，以及英文章节 3 篇。受邀撰写中国控制理论战略报告自适应控制部分。曾获中国自动化学会优秀论文、国际机器人与应用大会成就认可奖等国内外学术会议荣誉或奖励。所指导的学生多人次在科技竞赛中获得国际、国家或省部级奖励。

多尺度量子谐振子优化算法

西南民族大学计算机学院教授 王 鹏

近年来，云计算、大数据、智能计算等新概念层出不穷。多尺度量子谐振子优化算法属于人工智能领域。每一项新技术的出现，都与当前技术和产业情况相关，当云计算最早出现时，也是由于网络及计算机技术发展到一定阶段才提出的。每个新技术的出现，并不是凭空而来的，而是与生产力和生产关系相适应的。经历了云计算、大数据到人工智能的火热，这些技术均不是偶然出现的新技术，新技术的出现均与当前的技术发展趋势相关联。

人工智能火爆的标志是 2016 年 AlphaGo 战胜了围棋高手李世石，随后掀起了全世界研究人工智能的热潮，但人工智能的火爆并不是单纯由某一事件而引起的，而是与技术发展到一定阶段有关系。当技术发展到一定阶段后，人工智能的火热便是必然出现的一个结果。云计算、大数据技术的积累都为人工智能的发展提供了很好的基础。

多尺度量子谐振子优化算法是我们团队组经过多年研究提出来的。

一、人工智能的历史

人工智能的发展经历了很多的起伏，最早的萌芽是从 1950 年“图灵测试”的提出开始的。到了 1956 年，人工智能界召开了达特茅斯会议，会议进行了两个月，提出了人工智能的概念。这次会议的召集者是麦卡锡，麦卡锡是一位非常著名的人工智能学者，被称为“人工智能之父”，也被称为“云计算之父”。麦卡锡早在 1965 年左右便提出了“计算是可以作为资源提供给用户使用”的概念，后来，他在《人工智能预言历史》一书明确采用了大数据的概念，他是非常有远见的，很早便对云计算、大数据技术提出了极具远见的预言。

人工智能的发展并不是一帆风顺的，在 1956 年达特茅斯会议之后，人工

智能的发展经历了高潮，也有低落。2016 年 AlphaGo 战胜人类最强的围棋棋手，将深度学习的人工智能技术重新推向了风口，使人工智能成为产业界的下一个重要风口。至此，云计算、大数据、人工智能相互融合，成为未来几十年产业发展的重要领域。

人工智能的提出涉及一个很重要的基础。当前，计算技术的发展与大数据技术的发展是密不可分的，AlphaGo 能战胜人类与其身后具有强大的并行计算机是有关系的。同时，大数据的出现使我们储备了大量知识，计算机能力及大数据的发展为人工智能再一次繁荣提供了很好的基础。因此，生产力必须适应生产关系，人工智能的火热并不是偶然的，是生产力发展到一定阶段后必然产生的，而 AlphaGo 仅仅是触发产业发展的一个契机。

二、量子力学与人工智能

量子力学是在 20 世纪初由一群世界上最聪明的科学家共同努力完成的，我们一直相信，由这些人构建起来的量子体系，定会对计算机科学起到很好的启发作用，量子力学终结了我们认为的世界是确定性的观点，从而使长久以来确定性的世界观完全崩塌。

通常，很多人认为量子力学与自己无关。站在历史的角度看，很多计算机科学家均受到量子力学的影响，目前的计算机之父冯·诺依曼写出了《量子力学的教学基础》，1931 年图灵认真研读了这本书，而且早在 1929 年图灵还着迷于天文学家爱丁顿所著的《物理世界的自然》一书，爱丁顿也认为“大脑也是由原子、电子组成的，量子物理或许能为人类意识、思维提供产生的机会和空间。”可以看出，这些科学家，尤其是计算机科学家，很早便认识到量子力学可能对自然科学提供帮助。

量子力学终结了确定性的世界观，传统的物理学认为，电子是按一定的轨道来运行的。按照电动力学的原理，如果电子按一定的轨道运行，将会在切线方向上辐射出光子，由于电子能量的降低电子将迅速坠落到原子核之中。只有通过量子力学将电子以概率的形式进行解释，电子在核外是以电子云的形式存在的，而不是通常认为的轨道形式，这是量子力学对物质世界的新认识，认为物质世界是由概率所支配的。因此，在设计新算法时充分利用这一

观点，包括量子波函数的观点来构造新的算法。

人工智能中的智能算法如遗传算法、粒子群算法、蚁群算法、模拟退火算法、膜计算等算法都是在向自然界学习，但很多算法都存在一定的弊端。一方面，很多算法没有完整的数学描述；另一方面，不少优化算法都是学习自然界的表面运动现象。基于量子理论的算法为我们提供了一个学习自然界本质规律的方法，通过这种方法构造的算法有可能真正反映自然界本质的运行规律。因此，我们一直以来坚持使用量子力学的基本原理构造新的算法，通过利用量子力学的数学框架为构造新的算法提供了充分的理论支持。

（一）量子谐振子算法提出已有七年时间

2010 年提出了该算法的雏形，当时称为模拟谐振子算法（SHOA）。从 2010 年一直持续到 2016 年。2013 年，将 MQHOA 算法的基本物理模型完整提出，将算法正式定名为多尺度量子谐振子算法（MQHOA）。随后，在该算法的理论和应用方面做了一些工作。特别是在 2016 年，将算法进行很大的改进，使算法能快速处理超高维的函数优化问题，对一些复杂高维函数进行优化处理，目前正在向上万维的复杂函数进行优化。

量子谐振子算法有一个很重要的等效关系，将优化问题的目标函数转变为求薛定谔方程中约束态下的基态波函数，具体如下：

$$\left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right) \psi(x) = E\psi(x)$$

$$V(x) = f(x)$$

但不幸的是，薛定谔方程的求解很困难，因此，采用了近似逼近，如下式所示，由此，任意目标函数在极值附近的泰勒二级近似就是谐振子势，对复杂目标函数进行逼近，用谐振子式代入薛定谔方程中。

$$V(x) = f(x)$$

$$= f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \dots$$

$$\left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + \frac{1}{2} kx^2 \right) \psi(x) = E\psi(x)$$

算法依据波函数图进行构造，在 高能态时要经过多个中间能级（亚稳态）才能达到量最低态（基态），其中，亚稳态对应于目标函数的局部最优解。

(二) MQHOA 算法的基本流程

整个流程包含 3 个过程：能级稳定过程、能级下降过程、尺度下降过程。算法的整个运行过程便是在 k 个采样的驱动下，不断对新的采样进行比对。

图 1 描述了算法的整个收敛过程。算法分为横向收敛和纵向收敛，横向是高能态逐步向基态收敛，一旦完成到基态的收敛过程，便将尺度减小一半，进入更小尺度的迭代，直到最后的尺度满足预先设定的大小，其中存在多尺度的迭代。

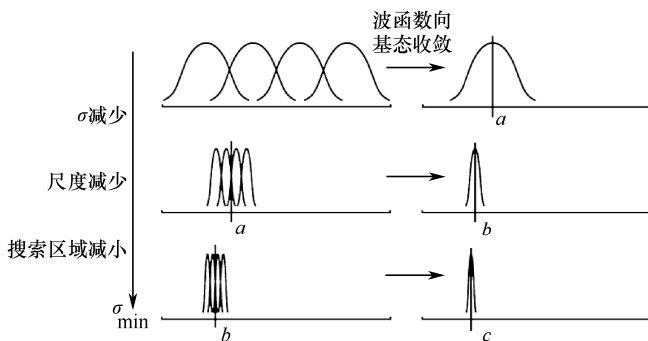


图 1 算法收敛过程

(三) MQHOA 算法的优点

(1) 其他很多的优化算法需不断调整参数，MQHOA 仅有一个需要主观控制的参数，便于我们调整算法的性能。

(2) MQHOA 算法具有良好的高维优化能力，目前已实现对一些复杂函数进行一千维以上的优化，用基本算法便可完成。

(3) 整个算法的迭代过程很简单，整个迭代过程就是不断进行采样，取代差解的过程。

MQHOA 算法的整个流程比较容易理解，但物理含义比较深刻，不易理解。对于从事应用工作的用户而言，只需掌握其流程即可。

MQHOA 算法还定义为波函数的概念，波函数便是将 k 个采样区域进行叠加。波函数有重要的物理含义：在量子力学中，是粒子出现的概率；在算法中，是最优解当前的分布。波函数会随着迭代过程发生变化，直到波函数

收敛到最优解附近。波函数在不同尺度、不同能级迭代时有着非常明显的逐步收敛过程，刚开始是在局部最优区域收敛，最终会在最优解附近。波函数在算法中的概念与量子力学中的概念是很相似的，都在描述物质出现的概率。整个算法的目标是将波函数的概率分布逐步向全局的最优解概率集中，以最大的概率采样到最优解。

在研究过程中出现了一个新的规律：算法的测不准原理。这个原理是在1927年由海森堡提出来的，测不准原理是说在物理世界一个粒子的位置和速度是不能同时被测定的。量子力学中有一个定理：如果两个物理量所对应的算符的对易式不等于零，其所定义的物理量便无法被同时精准测定。我们采用这一定理证明了算法的测不准关系。

优化算法的测不准关系具体是指，优化算法的全局搜索能力与局部搜索能力是不能被同时获得的。模仿量子力学的算符方法，定义为全局搜索算符和局部搜索算符，全局搜索的含义便是发现所有的全局最优解，为获得所有的全局最优解，在数学上的操作是对目标函数进行求导，所以，我们将全局搜索算符定义为对目标函数的求导。局部搜索是精确获得最优解的位置，将自变量的位置作为算符，按照量子力学中的定义式作用在函数上，最终证明了全局搜索和局部搜索是不能同时获得的。优化算法的测不准原理反映了算法的多尺度问题，很多智能优化算法均不同程度地使用了测不准原理，我们不能指望只用一个尺度让一个算法同时获得良好的全局搜索和局部搜索性能，这也为研究智能优化算法提供了非常有力的理论工具。

人工智能要解决的一个工作便是如何在一个巨大的解空间进行搜索，因此，需具有极强的搜索能力。所有的人工智能算法都在试图用最快的速度完成对解空间的快速搜索，当我们面对指数级巨大的解空间时，需要利用算法的隐含并行性。算法的隐含并行性指的并不是在并行计算机上运行的能力，目前对这方面的研究比较少。1976年图灵奖获得者 Robin 认为“应该放弃的只是以完全确定的方式获得结果，这种结果可能出错，然而出错的可能性微乎其微，也就是说可以把概率算法用到这类问题中来”。因此，产生了一个初步的认识，即不确定性造就了隐含并行性。

大多数的人工智能算法采用了概率方法，包括 MQHOA 算法、波函数。正是由于我们降低了对求解精度的要求，使我们获得了快速的搜索能力。

目前的 MQHOA 算法已成功地在很多领域得到了较好的应用,包括函数优化、数据挖掘等领域,算法是具有一定的生命力的。但同时,算法中还存在改进的空间,希望其具有更好的并行性,解决更复杂的优化问题。

关于云计算和大数据的教学,希望将并行计算技术作为一门基础课程。对理解云计算和大数据的体系结构具有很大的好处,对培养学生在分布式环境下的思维能力也有着很好的帮助。

作者简介

王鹏:教授,博士生导师,计算机科学博士后、金融工程博士后,四川省学术和技术带头人后备人选,广东省高教学会高职高专云计算与大数据教学专业委员会理事长,中国电子学会云计算专业委员会委员,中国计算机学会高性能计算专业委员会委员,四川省计算机学会理事,四川省计算机学会高性能计算专业委员会副主任委员,四川省政务云建设专家咨询组成员,2007—2008 年度成都市“一专多能”优秀青年教师,成都市软件行业协会理事,成都市科技攻关计划项目评审专家。曾任中国电子科技集团 38 所无人机地面数据系统主持设计师,作为中组部“西部之光”访问学者赴中国科学院高能物理研究所从事云计算大数据处理研究,挂职汕尾职业技术学院副院长。迈普通信技术股份有限公司云计算首席专家,《成都信息工程学院学报》编委,汕尾市科技顾问团首席顾问,广东工业大学兼职研究生导师,武汉职业技术学院、广州科技贸易职业学院兼职教授,2012 年成都市物联网产业领军人物。

深度学习的最新进展

浙江工业大学计算机科学与技术学院 王万良

一、深度学习的兴起

众所周知，深度学习是相对浅层学习而言的，已有的浅层学习包括神经网络 ANN、SVM、Boosting、最大熵方法等，其主要局限性在于深度的有限性，表示的学习能力也是有限的，如果要有一定深度，便会产生一系列问题，尤其是反馈到最后的误差很小，对神经网络参数的调整能力变得很弱。因此，引进了深度学习来克服这一系列的问题。

（一）深度学习的机理

深度学习主要解决的便是特征表达问题，由于一般的学习方法中的特征提取均为人工手动进行的，而深度学习是直接面向数据的，特征表达由学习自动完成，包括预处理、特征提取、特征选择。最后与其他学习算法相同，能进行机器学习算法，包括推理、预测、识别等。因此，深度学习有一个良好的特征表达，这也是深度学习识别成功的关键。相对于传统的做法，对特征可自动进行分类。深度学习的机理是基于人对视觉信息的了解。

1981 年，诺贝尔医学奖获得者美国神经生物学家 David Hubel 和 Torsten Wiesel 发现：人的视觉系统的信息处理不是整体处理的，而是分级处理的。高层的特征是底层特征的组合，从低层到高层的特征表示越来越抽象，越来越能表现语义或意图，这便是深度学习的思想。抽象层面越高，存在的可能猜测就越少，便越利于分类。

（二）深度学习的推动

2006 年，加拿大多伦多大学教授 Geoffrey Hinton 和他的学生在 *Science* 上发表的文章掀起了深度学习的浪潮，即提出理论。

2012 年, Hinton 组参加计算机视觉系统识别项目, ImageNet 使用 CNN 模型以超过第二名 10 个百分点的成绩夺取当年的竞赛冠军, 并引起了广泛的注意。

尤其是伴随着未来云计算、大数据时代的到来, 计算机能力的大幅提升, 使得深度学习模型在计算机视觉、自然语言处理及语音识别等众多领域都取得了较大的成功。有专家预测, 深度学习将在信息检索上取得重大突破。

(三) 深度学习的基本思想

假设系统 S 有 n 层 (S_1, \dots, S_n), 它的输入是 I , 输出是 O , 表示为 $I \Rightarrow S_1 \Rightarrow S_2 \Rightarrow \dots \Rightarrow S_n \Rightarrow O$ 。如果调整系统中参数, 使得它的输出 O 等于输出 I , 那么就可以自动获得输入 I 的一系列层次特征, 即 S_1, \dots, S_n 。通过这种方式, 便可以实现对输入信息进行分级表达。其优点为可通过学习一种深层非线性神经网络结构, 实现复杂函数的逼近, 表征输入数据分布式表示。

目前, 深度学习不仅在学术界, 而且在产业界也受到重视, 这是至此深度学习最重要的一个特征。在学术界与深度学习相关的关键词如计算机视觉、自然语言处理、机器学习、顶级期刊会议、公开课程代码等; 但推动机器学习发展的一个重要的推手还是产业, 如 Google Brain、21 世纪计算大会、百度 IDL、“中国大脑”、刷脸支付等。

二、重要模型

(一) 受限玻尔兹曼机

玻尔兹曼机是一种随机的递归神经网络, 由二值神经元构成, 每个神经元只取 0 和 1 两种状态。然而, 即使使用模拟退火算法, 这个网络的学习过程也非常慢。Hinton 在原来的基础上去掉了玻尔兹曼机同层之间的连接, 从而大大提高了学习效率。同时还有深度玻尔兹曼机、深度置信网络等, 这些网络主要是将一些相互间的连接去掉。因此, 它是由多层 RBM 堆叠而成的, 神经元可以分成显性神经元和隐性神经元, 显性神经元用于接受输入, 隐性神经元用以提取特征。

（二）卷积神经网络

深度学习的核心便是卷积，其中有很多参数，使得训练变得非常复杂，根据数据机理进行简化，尤其是采用了局部感受野、权值共享，以及时间或空间将采样的结构思想，使得网络中自由训练参数的个数大大减少，降低了网络参数选择的复杂度，但其基本原理仍为基于人或动物的视觉特征。

其中一个很重要的技术便是局部连接，全连接神经网络中的每个神经元的图像数据均有连接，其权值很多，根据视觉机理，眼睛所看到的不是全部，而仅为图像的某一部分，对每一部分用神经元连接，需要训练的权值便减少了很多。但仍有很多。根据视觉机理，每个局部的感受都是相同的。因此，可将每个神经元与局部图像连接时使用相同的权值，权值便大幅度下降，以及权值共享。

（三）自动编码器

自动编码器（Auto-Encoder，AE）是一种尽可能复现输入信号的神经网络。我们构造一个神经网络，如果调整神经网络的权值使得输出与输入相同，中间的神经元表示的是对原来信号的一种压缩和编码，AE 能找到可以代表原信息的主要成分，是一种非监督式的学习方法。

为提高编码器的性能，又提出了自动编码器的进步改进——稀疏自动编码器。稀疏自动编码器（Sparse Auto Encoder）是在自动编码器的基础上加上稀疏性约束，即约束每一层中的大部分节点都要为 0，只有少数不为 0。限制每一次得到的特征表达尽量稀疏，这样便可以简洁地表示原信息的主要成分。

降噪自动编码器（Denoising AutoEncoders，DA）是在自动编码器的基础上，在训练数据中加入噪声，让自动编码器学习去除这种噪声而获得实际输入。这就迫使编码器去学习输入信号的更加鲁棒的表达，这也是它的泛化能力比一般编码器强的原因。

（四）生成对抗网络

生成对抗网络的基本思想为在一个神经网络中包含了一对相互对抗的模型：一个是生成网络 G，用于逼近真实数据分布；另一个是判别网络 D，辨别样本来自真实数据集还是生成网络 G。两者均可以是非线性函数，如多层感知

机。G 的目的是使生成的数据能以假乱真，而 D 的目的则是正确区分真假数据。D 的存在使得 G 无须真实数据的先验知识或复杂建模也能学习逼近真实数据，当 D 无法判断 G 的生成数据时，G 和 D 达到纳什均衡。

三、深度学习的应用

深度学习的应用具体包括大数据分析、自然语言处理、计算机视觉、语音识别与合成等领域，但不局限于这些方面。如语音识别合成、计算机视觉、大数据分析、自然语言处理等。大数据不是一个新的内容，数据处理技术很早就存在，人们利用大数据技术获得了很多信息。但目前的数据很多很复杂，因此，目前的大数据问题更突出、更难用、更有用。而深度学习也是大数据分析的一个主要工具。

(1) 更突出。物联网的发展使数据越来越大，越来越复杂。数据量在 2010 年，全球便进入 ZB 年代（十万亿亿字节），2011 年数据进入 1.8ZB，2020 年物联网数据增量达到 40ZB 左右。直观来看，近两年产生的数据总量相当于人类有史以来的总和，因此，数据量大更加突出。

(2) 更难用。原先的机理分析、统计不再适用，更多地依靠相关性分析。舍恩伯格的《大数据时代》一书中说，我们没有必要非得知道现象背后的原因，而是要让数据自己发声。数据的相关关系能够帮助我们更好地了解这个世界。

(3) 更有用。是关系国计民生的内容。2015 年的悉尼思想领袖峰会上，物联网之父凯文·艾希顿说：“那些智能酒瓶、智能比基尼、智能水杯什么的，都是渣渣。多做一些如城市大数据的智慧服务、工业大数据等。”

（一）图像分类

在每个测试图像下写上正确的标签进行分类，显著的例子便是大规模视觉识别挑战赛，在此过程中首次应用卷积神经网络取得最好的效果，识别量大幅度降低，在未来的工作中始终保持领先地位。

（二）视频跟踪

利用深度学习跟踪人物的轨迹，如跟踪人脸、跟踪汽车均取得了很好的效果。

（三）生成对抗网络应用

如从文字描述生成图片、视频分析、将 2D 图像映射为 3D 形状、矢量空间运算等。近期的研究是将生成对抗网络用于医药研究领域。莫斯科物理科技学院（MIPT）首次将 GAN 应用在研发具有特定医疗属性的药物（如抗癌药物等）。

四、展望

几个重要的研究方向：大数据深度学习、认知神经网络、复杂神经网络实现、无标签数据的特征学习。总之，人工智能的发展已不是云遮雾障，而是迎来产业发展的黄金期，前途十分光明。

作者简介

王万良：现为国家级教学名师，享受国务院政府特殊津贴人员。入选国家首批“万人计划”教学名师，入选浙江省跨世纪学术带头人，浙江省高校中青年学科带头人。现任浙江工业大学计算机科学与技术学院院长，软件学院院长，控制理论与控制工程博士学位点负责人，控制理论与控制工程专业博士生导师，技术经济及管理专业博士生导师，国家精品课程《自动控制原理》负责人，校优秀课程《人工智能导论》负责人。杭州市计算机学会理事长。

大数据时代的编程语言

中国人民大学信息资源管理学院副教授 夏 天

编程语言是大数据分析处理的必备工具，但编程语言的种类和数量非常多。以个人所接触过的编程语言为例，在过去二十年的时间中，先后学过的语言有 Pascal、C/C++、Visual Basic、Delphi、ASP、PHP、Java、Python、R、Scala 等，除了常用的 Scala、Java 和 Python 外，业界还有 Ruby、Ada 等其他众多的优秀语言。面对这么多编程语言，我们该如何选择？下面就一起来看一下在大数据处理分析方面常用的语言及其特点，以方便不同专业背景和兴趣爱好的人员，能够结合自身特点选择合适的编程语言，方便大数据的挖掘分析。

一、大数据时代数据分析处理常用的编程语言有哪些

为了了解大数据时代数据分析处理的常用编程语言，下面基于 YouTube、Quora 和知乎三个重要的社交媒体平台，以编程语言和大数据作为检索条件，探究有哪些发现。

YouTube 检索结果的第一页，共有 20 条记录，在这 20 条记录中，Python 出现了 8 次，R 出现了 4 次，Java 出现了 3 次。由于开源所使用的 Hadoop 和 MapReduce 主要是由 Java 语言编写而成的，因此，如果标题中未直接提到某个具体的编程语言，但又涉及了 Hadoop 或 MapReduce，也可归为 Java 语言，那么，Java 总计出现了 6 次，除此之外，Scala 和 Ruby 各出现了一次。

Quora 的检索结果中主要涉及了 Python 和 R 语言，也涉及了 Java 和 Scala 的相关讨论，在大数据话题下关于 Python 和 R 语言的争论是比较热门的讨论内容。

知乎的检索结果主要涉及 Java、Python、Scala、R 语言、C 语言和 C++

语言，其中以 Java 和 Python 为主。

除了上述典型的 3 个社交媒体之外，还可考虑当前的大数据生态圈。当前的大数据生态圈主要以 Hadoop/MapReduce、Spark 和 Storm 为典型代表，而这 3 个框架无一例外，都是构建在 JVM（Java Virtual Machine）——Java 虚拟机上的，从这个角度来看，Java 语言和大数据的关联最为密切。除此之外，Spark 除了 Java 之外，还提供了 Scala、Python 和 R 语言的相关接口。

综合来看，目前大数据分析处理使用最为广泛的编程语言主要有 4 种：Java、Python、Scala 和 R 语言。

二、大数据领域四大编程语言的特点

世界上有一个著名的编程语言排行榜——TIOBE，可以反映编程语言的热门程度。在 2017 年 3 月的排行中，Java、Python、R 和 Scala 分别占据了第 1、第 5、第 13 和第 30 的位置。Java 在绝大多数时间中都占据了排行榜的首位，足见其本身的流行度。

（一）Java

Java 是目前编写大数据底层平台的主要语言，我们用的 Hadoop、Spark 和 Storm 主要由 Java 语言编写而成，可见 Java 语言在大数据方面的重要作用。

Java 最初是由 Sun 公司的詹姆斯·高斯林开发出来的。高斯林长期供职于 Sun 公司，但 Sun 公司在 2009 年被甲骨文收购，高斯林一直对甲骨文公司抱有不满意，并于 2010 年离开甲骨文加入谷歌，他在离开甲骨文时说：“我所说的都关乎细节与诚实，但吐露真相只会带来更多的坏处”“在 Sun 与甲骨文的并购会议上，到处是有关 Sun 和谷歌专利的争吵。甲骨文律师的眼睛闪闪发光。”关于 Java 的故事虽然充满了传奇，比如那些不可思议的成功、失之交臂的良机、纠缠不清的官司，但 Java 本身却成功地应用在网络计算、移动等各个领域。历史上从未有像 Java 这样可以如此广泛应用的语言和平台。

Java 语言如此成功与 Java 的特点是密切相关的。在 Java 之前，最流行的编程语言是 C 语言，但 C 语言的指针处理对普通程序员而言简直是梦魇，经常会带来内层泄露问题。因此，高斯林发明 Java 的一个重要原因便是为了解

决 C 语言的内存泄露问题和 C++兼容 C 语法而造成的一些历史遗留问题。Java 提出了引用类型，取消了 C 语言中特殊的指针语法，并通过垃圾自动回收机制，自动回收不再使用的对象所占据的内存空间。程序员只需创建对象，无须关心如何回收对象，这就使得程序员的犯错率大大降低，开发效率也随之提升。同时，Java 还提出了中间语言和虚拟机的概念，Java 程序会先编译成名为字节码的中间语言，再由运行在各操作系统上的 Java 虚拟机软件（JVM）在运行时解释和执行。这样做的好处是实现了当年 Java 的口号：一次编写、到处运行，使得企业可以自由选择操作系统和服务设备，保护了企业的软件投资。

我们直觉上认为，Java 编写的代码需要编译为字节码，再由虚拟机来解释执行，运行速度应该慢于直接编译为机器码的语言，如 C 语言。但人们对 Java 虚拟机做了大量的优化，使得普通程序员编写的 Java 程序远快于普通程序员编写的 C 语言程序，不同水平的 Java 程序员也能较好地进行团队协作，开发大型项目，变相降低了企业的人力成本。

Java 具有跨平台特性和开放特性，编写效率相对较高，属于强类型的静态语言，便于大型项目的组织管理和模块划分。因此，毕业于美国斯坦福大学的 Hadoop 之父 Doug Cutting 在编写 Hadoop 时选择了 Java 语言，这并非偶然。

对于 Java 语言的特点，人们比较认可的是：简单、面向对象、分布式、健壮、多线程、安全、可移植、动态等。当然，其中的简单、多线程是相对于当时的 C、C++而言的。总之，Java 构成了当前企业大数据计算的基石。

（二）Python

1989 年的圣诞节期间，在荷兰的阿姆斯特丹，年轻的 Guido van Rossum 为了打发圣诞节的无聊，决定开发一个新的脚本解释程序，由于他非常喜欢英国六人喜剧团体 Monty Python，因此以 Python 作为脚本名称。

Python 由 ABC 语言继承而来，非常适合非程序员来学习使用，普通人员在学习 Python 时入门很容易，刚刚提到 Java 是一门简单的语言，Python 则是“更加简单”的语言。我们在开始学习 Java 时，需要配置环境变量、安装基础开发环境、编译运行等，这些基础工作使得很多人抓狂，甚至“成功”地将很多人吓跑了。Python 则简单得多，下载安装之后，在命令行直接输入 python，

便可打开一个解释器，每执行一行代码即刻能看到输出结果。因此，初学者非常乐意将 Python 当作一个强大的科学计算器来使用，体验 Python 的强大功能。

Python 遵循了优雅、明确、简单的设计原则，一件事情会有很多种方法，其中会有相对较好的一种。Python 开发者的哲学是用一种方法，最好只用一种方法做一件事情，如果有多种选择，Python 开发者会拒绝花哨的语法，而是选择明确的、没有歧义的语法。

Python 强制使用空格作为逻辑代码块的隶属关系控制，强制程序员养成良好的编程习惯。Python 的解释器中会输出 Python 推荐的编写风格和准则，例如，“优美胜于丑陋，明了胜于晦涩……”。在 Python 显示器中输入 `import this`，便可看到这一系列准则的英文原文。

经过二十多年的发展，Python 已变得非常流行。从早期的各类系统管理任务和 Web 编程，到后来的科学计算和数据分析，Python 的应用都比较多。Python 的简洁、易读和可扩展性使得用 Python 做科学计算的研究机构日益增多，如卡内基梅隆大学的编程基础、MIT 的计算机科学基编程导论都采用了 Python 来教授这些课程。此外，众多开源的科学计算软件包也都提供了 Python 调用接口，如著名的图形库（OpenCV）、三维可视化库（VTK），在科学计算和数据分析方面也形成了较为统一的经典扩展库，如用于快速数组处理的 NumPy、数值运算专用的 SciPy、图形绘制的 Matplotlib、金融处理方面的 Pandas。同时，还有大量的经典图书采用了 Python 语言来讲解数据挖掘方面的相关理论，如《集体智慧编程》《社交媒体的数据挖掘与分析》等。

因此，Python 的特点可以总结为简单、优雅、易于扩展、有丰富的科学计算和数据分析扩展库，非常适合数据科学家来学习使用。

（三）R 语言

R 语言本质上是一款集统计分析和可视化于一体的免费的可跨平台运行的统计软件。由新西兰奥克兰大学的 Robert Gentleman 和 Ross Ihaka 在 20 世纪 90 年代初期共同发明，当时两个人教授一门初等的统计课程，为方便授课开发了这一语言，由于两个人的名字均以 R 开头，因此，被称为 R 语言。

R 语言可以看作大名鼎鼎的贝尔实验室所开发的 S 语言的一种实现。1975 年，贝尔实验室的统计研究部使用了一套文档齐全的扩充库来做统计研究，

我们称为 SCS。但 SCS 在做统计分析时需要大量的编程，有人认为这样太麻烦，于是贝尔实验室又开发了一套完整的高级语言系统，即 S 语言，用于交互。1993 年，S 语言的许可证被 MathSoft 公司买断了，引起了人们的担忧，而开源的 R 语言引起了人们的关注。1997 年，R 语言正式成为 GNU 项目，大量的优秀统计学家加入了 R 语言的开发行列。到今天，这一场开源和商业、开放和封闭之争算是尘埃落定，R 语言已成为当今最为流行的统计分析工具之一。

与 Python 类似，R 语言的使用也很简单，直接打开 R 语言的交互环境，输入 demo，便可以看到其中的一些例子。因此，初学者可以很快地获得直观的感受，增加学习的乐趣，对学习统计知识非常有帮助。

从 R 语言的发展历史来看，主要是为统计学家解决数据分析问题而开发的。R 语言的特点是擅长数据的统计分析。R 语言提供的算法几乎覆盖了整个统计领域的前沿算法，重复性的工作借助 R 语言强大的分析能力和排版能力，利用 Sweave 能得到很好的解决。R 语言本身属于统计编程类语言，受到其算法架构的通用性和速度、性能等方面的影响，最开始的设计完全是基于单线程和内存计算完成的。因此，在处理大规模数据时显得力不从心，好在 R 语言有一些优秀的扩展，能在一定程度上解决这些问题，如 SparkR、RHadoop 等。

（四）Scala 语言

Java 促进了今日信息技术的辉煌，Java 支撑了大量的企业级关键应用，但人们一直期待 Java 的重大改进和更新换代，使其吸收其他语言的长处，进一步提升生产效率。而 Scala 不仅在“更好的 Java”（Better Java）方面做得非常成功，而且在并发编程、大数据处理、科学计算方面都取得了不错的成果。

Scala 由瑞士联邦理工学院的 Martin Odersky 在 2003 年开始设计，在此之前，他因对 Java 的优化而闻名于世，Scala 是其又一成名之作。正如其名，Scala 本身是一门可扩展的语言，其中有静态运行、面向对象编程、函数式编程、类型推导、高阶函数等众多特性。Scala 语言非常经典，相同的功能用 Scala 实现，代码量能达到 Java 的 20% 左右。Scala 不仅可以做到“更好的 Java”，而且在数据统计分析方面，原先 Python 和 R 语言独具的一些统计分析模块，已经有越来越多的 Scala 实现，如 Breeze、Spark data frame、Zeppelin。以前的数据分析书籍大多通过 Python 语言进行讲解，现在也有许多用 Scala 进行

数据分析和机器学习的书籍，如 *Scala Data Analysis Cookbook*、*Scala for Machine Learning*。

Scala 可以无缝衔接原有的 Java 库，充分利用现有的 Java 资源，在大数据并行分析处理方面更是独具优势。因此，很多大数据平台都选择 Scala 作为首要的实现语言和 API 接口语言。

个人认为，Scala 最主要的特点是太灵活，以至于初学者的学习曲线比较陡峭，但 Scala 很好地将 Java 的面向对象的编程方法与函数式编程思想揉在一起，其中体现的现代化编程思想值得程序设计人员深入学习和体会。

这 4 种编程语言各有各的优点，但有人地方便有争论，争论焦点主要集中在 R 和 Python 的选择上。其实，通过这两种语言的官网介绍，便可以看出它们的定位并不完全一致，Python 是一种通用的编程语言，除数据分析外，在很多领域都有广泛的应用，R 语言的定位是“用于统计计算的免费软件环境”。

三、初学者如何选择大数据编程语言

工欲善其事，必先利其器，对于奋战在大数据处理和分析前沿的人员而言，手中应有一把使用自如的快刀，当面对上述 4 种各有特色的编程语言，到底该如何选择？

在 Quora 上，有人提到，对于数据工程推荐使用 Java 和 Scala，对于数据科学推荐使用 Python 和 R 语言，个人觉得很有道理。如果你是一个统计学家，并没有太多的计算机科学背景，R 语言应该是一个很不错的选择；如果接受过较为系统的程序设计训练，又想从事数据分析方面的工作，Python 应该是不错的选择；如果我们要做大数据产品、大数据具体项目，或者想深入了解现在大数据框架底层的运行机制，那学习 Java 是必不可少的基本要求；对于 Scala 语言，如果你是一个熟练的 Java 工程师，厌倦了 Java 的冗余烦琐，向往美好的函数式编程、向往 Python 的简洁优雅、向往 R 语言的强大统计功能，可以选择 Scala。Scala 还有一个“额外的好处”：用 Scala 写的代码初级程序员很难看懂，也不敢随意修改，便不用担心在团队协作时精心编写的代码被一些初级程序员改错了。

当然，个人认为，编程语言并没有高低贵贱和优劣之分，各有各的优点，如果你有兴趣和能力，不妨将各种语言都实践一遍，尤其是侧重于大数据工程的技术人员，在学好一门重量级的语言（如 Java 语言）的同时，掌握一门 Python 或 R 语言，对于开拓自己的视野、学习数据分析的理论、阅读相关的书籍都是大有裨益的。

作者简介

夏天：中国人民大学信息资源管理学院副教授，电子文件管理研究中心、数据工程与知识工程教育部重点实验室研究员；美国 IUB 大学访问学者。研究兴趣包括：现代信息检索、Web 数据挖掘、电子文件管理。在 ICWSM、ASIST、iConference、ACM Hypertext 等图书情报和 Web 挖掘重要国际会议发表论文 30 余篇；出版专著和教材各一部，参编 4 部；主持和参与国家、省部级等项目 10 余项；研发的 Web 内容定向采集、网页正文抽取、主题链接抽取、关键词抽取、层次语义路径生成等模块得到了较好的实际应用，系统解决了 Web 数据分析处理的基础问题，同时是 GitHub 上 Xsimilarity 和 WikIT 开源项目的创建者和主要贡献者。

人工智能时代的“盲点”——信息无障碍

浙江大学计算机科学与技术学院常务副院长 卜佳俊

一、背景与意义

众所周知，新一代人工智能规划已经由国务院正式印发，对于这个规划普遍的解读是蕴藏了很多的机会，也是中国人工智能领域弯道超车的机会。在规划中，针对技术和软件领域提出了大数据智能、跨媒体智能、群体智能、混合增强智能、自主智能系统五大方向，同时，提出了包括智慧城市、智慧医疗、智慧制造等多个行业和领域的实际应用。

从目前我们所处的实际环境来看，在很多领域已经诞生了众多的人工智能应用，如我们熟悉的计算机视觉、语音识别、自然语言处理等领域。

牛津大学和耶鲁大学的研究人员甚至列出了一张 AI 彻底替代人类的时间表，2018 年开始，会有各种各样的人类工作逐步被 AI 替代，预计到 2136 年，人类所有的工作都全部由 AI 控制实现自动化。

但正如人类的视觉盲点（视网膜上无感光细胞的部位称为盲点，盲点是视神经穿过的地方）一样，目前人工智能的研究可能也面临同样的“盲点”。根据最近 10 年的分析、了解与研究，我们认为信息无障碍正是人工智能领域的一个“盲点”，大多数的人重点关注如物联网、视觉、语音、自动驾驶等领域相关的应用，但在信息无障碍领域真正关注和应用人工智能的还相对欠缺。

什么是信息无障碍呢？信息无障碍是指任何人（无论健全人还是残疾人，无论年轻人还是老年人）在任何情况下都能平等、方便地理解、交互和利用信息。这个概念最早源于我们对特殊人群，特别是残疾人的关心，但实际上信息无障碍是对任何人都是有用的，如我们开车的时候、在国外听不懂对方语言的时候，等等。在我国，特殊人群的数量是巨大的，如果把老年人、残疾人都计算在内的话，占总人口的 20% 以上。

从国家战略需求的角度看,无论国家“十三五”规划提出的“确保全面小康”“一个人都不落下”,还是“一带一路”倡议,都对信息无障碍的理论和技術提出了更高的要求,如果没有这方面新理论、新模型、新方法的突破,特殊人群将无法享受信息社会带来的福利,也就无法同步小康,也不能更好地实现“一带一路”建议。同时,从近期国际发展趋势来看,不管是国际学术界,还是西方发达国家及联合国等国际组织等,都在这方面做了不懈的努力,我们国家从21世纪初开始跟踪这个领域,近十年来这个领域的研究工作和成果也在飞速发展。

综上所述,信息无障碍已成为全球各界关注的热点,服务特殊人群已成为当今社会的主流理念,而缺乏体系化的信息无障碍理论研究,无法解决特殊人群信息获取与交互的障碍,成为人工智能时代的“盲点”。

二、信息无障碍领域研究工作

人类在获取外界信息的过程中,视觉通道占80%以上,听觉通道占10%以上,其他的如嗅觉、触觉等仅占不到10%,因此,在信息获取方面,盲人和聋人面临的困难是最为突出的。人工智能技术飞速发展带来新机遇的同时,也给以残疾人代表的特殊人群带来了新的挑战,如打车软件、个性化推荐应用,如果没有很好地利用信息无障碍的技术和手段,将会给特殊人群带来更大的障碍。

下面以盲人为例,通过3张图片列举信息无障碍问题的3个实例。多媒体信息主要通过视觉和听觉两个通道传达给健全人,但盲人缺失了视觉通道,势必给视觉通道上的信息传播带来障碍,当然我们也不是没有办法,一种办法是将视觉信息转换成文本信息(我们称为替代文本),然后将文本信息转换成语音信息以便让盲人获取。但从目前的研究和实践来看,二者之间的转换,人工智能技术还未达到非常完善的程度,即便是已经形成了相应的替代文本并变成了声音,但听力通道能传达的信息速度要比视觉通道能传达的慢很多,因此,存在第二个问题,即带宽窄。第三个问题是实时困难。

简单总结一下上述三个问题,通道受阻会造成“语义鸿沟”的科学问题,带宽窄会造成“信息过载”的科学问题,实时困难会造成“实时高效”的科

学问题。因此，我们认为，信息无障碍的研究体系应该在人工智能、大数据及对特殊人群的特殊需求充分了解、调研和研究的基础上架构起来的，包括底层基础理论的研究、中间关键支撑技术的研究，以及上层的系统和相应的服务，当然贯穿始终的还要有一个统一的规范标准体系的支撑。

从“语义鸿沟”的角度出发，我们需要寻找一种数据的低维表达方式对多媒体数据进行分析，从而帮助理解多媒体的语义信息；从“信息过载”的角度，我们希望能有更好的文本摘要和个性化推荐等算法，从而帮助残疾人或特殊群体用户能在有限带宽内获得相应的信息；从“实时困难”的角度，针对健全人实时交互的辅助能达到计算快、可实时，对特殊人群主要是要进行多通道、可交互方面的研究。

针对上述科学问题列举以下 3 个研究的例子：①社交媒体推荐，互联网上有海量的有声资源，包括音乐，如果能对用户的收听和收藏行为有很好的了解，进行更好的个性化推荐是非常有意义的，我们通过超图建模音乐社区中的各类异构对象，包括用户、音乐、标签等，并提出了基于图传递的个性化音乐推荐算法，实现音乐的精准推荐，这个研究成果获多媒体领域最好的国际会议 ACM Multimedia 的最佳论文提名奖；②文档摘要，传统的信息摘要方法主要考虑覆盖原文中心思想和尽量减少冗余，部分算法还需人工干预，我们首次提出了基于信息重构的摘要方法，将重构误差最小的信息子集作为原信息的摘要，该研究成果获人工智能领域最好的国际会议 AAAI 的最佳论文奖，这是 AAAI 创办 30 多年来中国学者迄今唯一获奖的一次；③色彩转化，我们提出了一种新的把彩色图像变成灰度图像的方法，其结果比 Photoshop 的效果还要好，该研究被该领域最好的国际期刊 *IEEE TPAMI* 发表。

针对这些科学问题的研究及相应的成果，有 3 个具体的应用：①中国盲人数字图书馆。这是首个面向视力残疾人的国家级专门图书馆，利用文本摘要、个性化推荐等领域相关的模型、方法和算法实现了一系列的智能模块，目前已为 116 个国家的 300 多万人提供了盲用多媒体资源阅读和检索服务。②针对盲人的电视无障碍网络直播系统。该应用已经覆盖了 15 个国家的 50 万人。③面向所有残疾人的中国残疾人服务网。这是首家依托残疾人服务体系的大型公益门户，突破了以往残疾人服务只依赖于线下的局限，目前已向 1000 多万名用户提供了信息资源推荐服务。

三、信息无障碍领域国家标准与检测

在信息无障碍领域，我们能做的事情很多，特别是人工智能的新模型和新算法有很多地方可以应用，但如果没有标准规范及相应检测手段的支持，大家各自为政的话，最后造成的结果往往还是会有障碍的。

（一）信息无障碍领域国内外标准现状

目前 Web 内容的无障碍国际标准《Web 内容无障碍指南 2.0》来源于 W3C，美国、英国及西方很多发达国家，都是直接采用该标准或者在此标准基础上根据自己国家的实际情况定制相应的标准。我国也有自己的行业标准和国家标准，主要也是基于《Web 内容无障碍指南 2.0》的基本框架。目前我国正在制定最新的互联网内容无障碍的国家标准，预计 2018 年年底正式发布。当然，标准仅仅是一个方面，标准的推广和实施更为重要，因为即使有标准，如果大家不按照标准来实现，其结果对残疾人而言还是会有很大的障碍，这样的结果是没有意义的。以 Z 政府网站为例，截至 2015 年 7 月，政府网站共有 85890 个，如何推动这些网站的无障碍化，并给其他类型的网站做示范，这是非常有意义的。从实践的角度看，不仅是网站无障碍改造的推动难度较大（其实技术上的难度不大，主要是体制机制及其他非技术因素），而且改造完成后是否达标的检测也很难。为此我们专门研发了一个基于群智技术的网站无障碍合规检测系统。

（二）网站无障碍合规检测系统

之前网站无障碍合规检测主要以人工检测为主，从 2013 年开始自主研发了基于群智技术的网站无障碍合规检测系统。该系统能进行全网站检测，支持增量、采样、全站镜像的智能爬虫，能自动检测网页模板，并能自动生成检测报告并提供改进意见。为了更直观地表示网站无障碍合规监测的结果，我们将网站信息无障碍评估体系分为 100 分，其中可感知性、可操作性各占 38 分，可理解性和兼容性各占 12 分。

系统自投入使用以来，完成了 2013—2016 年中国政府网站和社会网站的

无障碍合规检测总计超过 3000 个网站。在 3000 个网站的检测过程中我们也发现了一些规律，例如，2015 年的中国社会网站检测结果显示，网站无障碍的平均分数不足 60 分，所以，总体而言中国目前网站内容无障碍的情况是不容乐观的。从技术的角度进一步分析，我们以检测残联系统的 132 个网站为例，共爬取了 83 万多个网页，自动检测点有 959 万多个，抽样后的人工检测点还有 25000 多项，可见检测的工作量是非常大的，需要大量利用大数据及相关的人工智能技术进行加速。

除了技术的攻关以外，大家联合在一起发声也是非常重要的。因此，从 2016 年 9 月开始，我们成立了信息无障碍技术标准联合工作小组，在中国残联和国标委的共同指导下，由中国残联信息中心、中国残疾人信息和无障碍技术研究中心作为共同组长单位，成员单位包括了行业内的一些协会、以 BAT 为代表的互联网企业，以及 IBM、科大讯飞等其他一些在无障碍领域做得比较好的公司。联合工作小组的目标是做信息无障碍国家标准总体框架体系。

四、建议

（一）国家层面

国家应尽快建立健全更系统的信息无障碍法律法规和标准规范体系。此外，整个国民教育体系对人工智能的“盲点”技术的关注和教育也是非常重要的，到目前为止，我国还没有一门课、一个专业、一个学院、一所学校专门将信息无障碍的理念、技术、标准规范体现在课堂教育中。其次，标准规范内容本身也是很重要的，要将现有的标准统一，甚至包括国际标准，这是因为互联网是全世界相通的，没有特区。

（二）开发人员

我们知道，信息无障碍的这些技术往往是融在产品 and 系统中的，如果在产品和系统设计的过程中能充分考虑所有人的需求，那么在技术实现时将信息无障碍的问题解决，后续便不会有更多的问题，甚至也不涉及所谓无障碍改造的问题。

（三）科研人员

希望有更多人工智能领域的专家关注信息无障碍领域的研究与发展，人工智能对信息交互能力的提升是适用于每个人的，健全人可以获取大部分信息，信息获取提升的幅度相对较小，但对特殊人群而言，利用人工智能技术对其信息获取能力的提升就会非常大，有时候甚至是 0 到 1 的变化。如果我们能在这一方面做更多的工作，一定会方便所有人获取更多自己需要的信息。

现在已是全人类的 AI 时代。人工智能的发展能有效推动信息无障碍领域相关问题的解决，这些问题的解决，对每个人而言，不论健全的，还是残疾的；不论青年的，还是老年的，都能获取自己需要的信息，在获取信息的基础上，能方便自己的工作、学习、生活和社交。同时，人工智能也会缩小健全人与残疾人之间的差距，但这是一个双刃剑，如果事情做得好，差距会越来越小，如果不被关注，差距会越来越大。因为健全人能用一个打车软件方便地打到车，如果这个打车软件对残疾人有障碍，那么残疾人能打到车的概率会越来越小，离主流社会的距离会越来越大。我们在应用人工智能及相关的技术时，如果不能把无障碍的事情做好，“盲点”势必会越来越大，信息鸿沟会越来越大，人与人之间的差距也会越来越大。

乐观地看，目前所有人都能享受到人工智能带来的便利机会，更多的国家、更多的科研人员、更多的开发人员已经开始关注和了解信息无障碍这一领域，不论从基础理论的角度、关键支撑技术的角度，还是从系统服务的角度，所以，我们相信在不久的将来，人工智能的“盲点”将不复存在。

作者简介


卜佳俊：浙江大学软件学院常务副院长，浙江大学计算机科学与技术学院教授。浙江大学计算机学院软件研究所副所长，中国残疾人信息和无障碍技术研究中心副主任，浙江省服务机器人重点实验室副主任。2005 年度浙江省“新世纪 151 人才工程”第三层次培养人员，2008 年度浙江省“新世纪 151

人才工程”第二层次培养人员，2012 年度浙江省“新世纪 151 人才工程”第一层次培养人员，2009 年度教育部“新世纪优秀人才支持计划”入选者。现任中国计算机学会杭州分部主席，中国计算机学会青年计算机科技论坛（CCF YOCSEF）主席，曾任 YOCSEF 杭州学术委员会学术秘书、副主席、主席等职。主要研究方向有嵌入式软件与传感网、数据挖掘与智能检索、媒体计算等。



应 用 篇

融合先行，释放价值



工业/物联网大数据

复旦大学计算机科学技术学院副院长 汪 卫

一、工业大数据的产生

目前，广为熟知的是互联网方面的大数据，实际上在工业领域中也存在很多大数据来源。由于工业中有大量的传感器，且 24 小时不停地工作，从很多统计来看，工业领域的大数据规模比互联网的大数据规模还要大。

在当前以第四次工业革命为代表的工业发展趋势中，如德国的工业 4.0、美国的工业互联网、我国提出的《智能制造 2025》都将数据分析能力作为主要的一部分，这也使得大数据成为工业领域当前的一项重要支撑技术，甚至像机翼等传统的制造业企业也号称逐渐向软件企业转变。在传统的工业领域中实际已经做了大量的信息化建设，包括建立 PRM、ERP、CRM 系统，这些系统对整个工业生产过程中的数据进行了有效的管理，当然，近年来互联网中大量的数据分析，使得传统企业对互联网数据的获取能力有了很大的提升。今天的工业大数据中所带来的与以前的不同之处在于，由于物联网技术应用的广泛和对行业的渗透，很多工业领域在各种设备和产品中安装了大量的传感器来获取信息。正是由于大量的机器、物联网产生的数据，为工业大数据带来了新的机遇。因此，工业大数据的产生实际是由于机器设备上各种传感器产生的海量数据，出于对这些数据的分析、存储需求，便诞生了工业大数据这一领域。

工业大数据的特点如下：产生了很多时间序列的数据，而且位置轨迹数据在工业大数据中发挥着很大的作用。工业数据中，在传感器数据的采样过程做非常密集的采样，这就使得工业大数据中的数据规模非常庞大，由此带来了存储分析上的技术产业。

二、工业大数据中的关键问题

从技术上来看,这种机器物联网数据带来了3个方面的技术挑战:

第一方面,存储规模和处理能力。如机器会每天24小时以很高的速度产生数据,这就需要我们面向流式计算的大规模分布式计算平台,承接由机器物联网产生的数据。同时,也需要具备海量数据的存储和高效查询能力,虽然原先已有海量数据库存储大量的数据,但在机器物联网中生成的数据类型与传统的数据库有一定的差别,如大量的区别数据,如何对这些数据进行有效的查询和存储是一个很重要的问题。

第二方面,在于对数据的深度分析需求。比如在序列数据中,单纯地看一个数据点的信息,并不能很好地反映数据的整体语义。因此,在序列中需要对一段序列进行语义上的分析,这样才能获得序列中真正蕴含的语义信息,这就需要有很强的事件发现能力。同时,在物联网社会中,我们针对一个设备或装备所能安装的传感器是非常多的,这就使得我们在分析时不是单道的时间序列,而是多道的时间序列做共同分析,甚至对其关系进行深入分析。因此,分析方式也有一定的差别。整个时间序列数据的存储方式、融合方式,由于其数据类型的独特性,为我们带来了许多挑战,传统的分析方法对于时间序列及内容的查询和分析是不足的。我们认为,工业大数据中的核心技术包括管理技术,主要是时空数据库,由于数据产生速度很快,这是一种紧密进行的数据库类型,而且时空数据的索引,尤其是序列数据的索引是工业大数据中很重要的内容。由于数据是机器产生的,所以数据质量应该是非常好的,但恰恰相反,在工业数据中的数据质量也是很重要的内容。

第三方面,工业数据中的数据并不是一个简单的区别数据,需要和整个装备与环境紧密结合。因此,工业数据中的建模也是一个新的挑战。工业数据中的处理平台和分析平台和原来相比也有很大的差别,比如,工业数据中更多强调并发的处理性能,以及分布式的处理能力。此外,对工业数据产生的大量原始数据语义分析也是工业数据中很难的问题,特别是机器构造原理与表现出来的现象之间的关联,这也是工业大数据分析所面临的一个问题。它不是一个单一数据类型的分析,还需要考虑基于这种设备的构造原理和模

块结构，基于构造原理和模块结构对数据进行分析，这是一个很重要的问题。

三、案例介绍

（一）大型车辆装备的监控

大型车辆装备的监控主要是国内的一个重型工程车辆生产厂商，对其所生产的各种车辆装备产品进行全面的数据收集与监控，通过对这些数据进行分析，可以实现对车辆装备的全面监控，如对车辆状态和车主状态的掌握。在重型车辆领域中，大部分车辆是以金融贷款的方式来获得的，金融部门需要对这些车辆的工作情况、位置进行监控，通过对车辆各种传感器的分析，便可以获取这些信息。还可以做大量关于备件计划的准备，如通过对车辆传感器的分析，可以得知车辆的运行状态。例如，可以通过对这些车辆的位置分析，得知这些车辆是否参加某些国家的重点工程，这些重点工程都在什么地方，对其地理环境进行分析，推测出机器设备中哪些部件会发生损坏。由于车上的部件也装有传感器，可以判断其是否有发生损坏的可能性。例如，曾经分析发现一大批产品在沿海地区施工，针对东南沿海高热的特点，第一时间对车辆部件进行准备，及时送到现场，以确保工程的顺利进行。通过这些分析，还可对车辆的设计进行改造，对车辆的各种故障发生及互相之间的关联分析，可以在设计过程中进行改进。

在机器数据分析时，任何一个机器装备都有其正常状态。当分析机器装备是否异常时，有一个基线模型，通过与基线模型比较，便可得知设备处于正常状态还是异常状态。因此，在机器数据分析中，很重要的一条是理解数据的来源，即某个传感器产生的数据符合什么变化模型，通过对实测数据和基准数据的比较，获得各种各样的信息，这里存在对领域数据理解的问题。

同时，在车辆数据的分析应用中，数据质量需要解决一个问题。由于车辆的工作环境较差，比如因位置原因导致数据传输丢失，没有通信网络数据便没有传到服务器中。由于所处环境信号不好，传输的数据有时会有异常值，有些异常值是由干扰产生的。有多个传感器获取数据，但它们的工作时间与获取数据的时间是不匹配的，大大影响了对装备的了解，如电压值和电流值的测试时间差。在记录压缩的过程中，也会丢失一些信息，因为有时会采用

有损压缩的方法。

（二）桥梁健康监测

上海地区水网密布，这里各种各样的桥梁达 3 万多个，还包括很多高架桥，这些桥梁的结构建造直接威胁了人们的生命安全。因此，对桥梁结构进行了监控，安装了大量的传感器，分析其是否安全。

其中，也遇到了一些关键技术，如事件识别，在桥梁上会装很多视频监控，但很多信息是无法获取的，如地震、桥墩撞击、爆破事件、台风、集装箱卡车经过等，这些信息对于桥梁的养护是很关键的，这就需要通过桥梁中的传感器来获取这些事件。不同的事件产生的信号是不同的，因此，如何通过不同的信号来分析不同的事件是目前很重要的一项工作。以重车经过为例，在桥面上不同的位置安装传感器，通过不同位置传感器的波形变化来判断是否有重车经过。但通过不同传感器感受到震动的时间差异，判断具体的事件信息。

另外，由于存在大量的传感器，每个传感器都在产生数据，传感器在分布式平台上生成处理时，会发现由于传感器的不同，需要每个传感器编写一个程序，这是很困难的。特别是在分布式平台上的程序编写，因此，开发了一个自动查询生成引擎，定义了一个数据访问语言，自动生成查询过程，进而提高了对数据的处理效率。

（三）卫星测试数据分析

随着国家航天技术的发展，卫星发射的速度越来越快。这就需要减少测试周期，提高卫星发射效率。这里会得到大量类似于电流的信号，这些信号以波形的方式出现，针对这种问题，研究了波形数据模型，建立一种模型，分析其本身处于什么数据状态。由于对数据进行了事件识别与分析，可以发现不同的数据项、数据序列之间的关联，如电压上升时电流值的变化分析。当发现电压/电流值不符合规律时，则判定为不符合规律的故障。

另外，还做了大量关于时间序列的相似性查询工作，系统中存在历史上积累的大量电压电流波动值，有时在故障分析时需要将这些信息对比。为此提出了一种结构，对波形变化的趋势进行特征描述，以便于用户直接在

数据库中查找。此试验已有了很好的效果。

作者简介

汪卫：复旦大学计算机科学技术学院教授、副院长，毕业于山东大学计算机系。目前担任中国计算机学会数据库专业委员会委员，上海市计算机学会理事，数据库专业委员会副主任，并担任 ICDM、SIAM DM、CIKM 等重要国际学术会议的程序委员。长期从事数据库与数据挖掘领域的研究和开发工作。

大数据语义分析与应用实践

北京理工大学计算机学院副教授 张华平

一、语义：比 AlphaGo 更难的事

大数据的语义分析对人类语言的理解难度远远难于 AlphaGo, 如就同样的文字而言：“谁都打不过”，意思上是可以完全相反的。由此可以看出，语义理解的困难所在。例如，“WE DO CHICKEN RIGHT”，真正按照文字的字面理解，这里涉及很多语言歧义。

图 1 中构建了自然语言、思维与客观世界的三角关系。可以看到，自然语言是人类理解客观世界的必要通道，几乎也是唯一的通道。

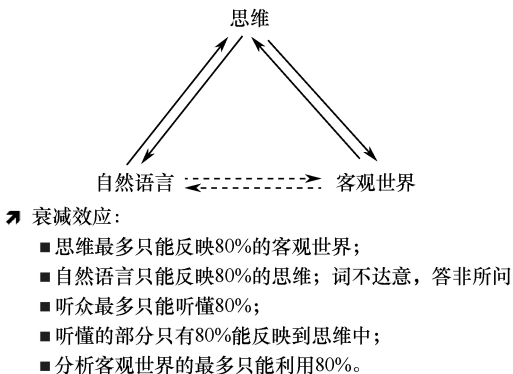


图 1 客观世界、思维与自然语言

二、文本大数据挖掘关键技术

从图 2 中可以看出，大数据更大意义上是非结构化内容理解。具体而言，结构化的大数据分析是利用传统的数据库，包括 SPSS、IBM 的 DB2 等这些工具可以很好地解决。但非结构化的内容理解还远远无法做到。

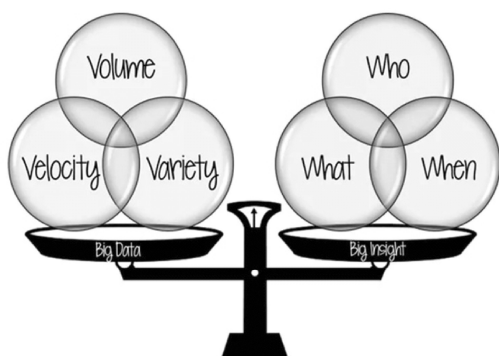


图2 大数据非结构化内容理解

实验室的主要研究内容包括 NLPiR 大数据语义挖掘、JZSearch 精准搜索引擎、知识本体构建与知识管理。

实验室历时 15 年开发了一个 NLPiR 的大数据语义分析平台。其核心功能包括以下几个方面。

搜索类：全文精准检索。

语言类：新词发现，分词标注，统计分析与术语翻译；关键词提取。

文档类：文本聚类及热点分析；分类过滤；自动摘要；文档去重；情感分析。

除此之外，还有一个 NLPiR 在线演示平台，下面将对在线演示平台的几个关键功能逐一进行介绍。

图 3 展示了一个基于在线演示平台做的，被称为新词发现的技术。它可以对一批语料自动计算数据中出现的新的词汇，如认沽权证、金融衍生产品等。新词发现结果包含几个参数：词语、词性（一般是名词）、权重（通过信息熵来计算该词对一批语料的重要性）和词频，这里的词频排第一的并不是最高的，因此不适用所谓的高频分析。另外，通过新词发现可以大量识别网络中出现的新的语言及专业词汇，这种方法分别在电力、医院做过实验，证明可以非常精准地识别各种专业的词汇，如药物名称、医学典籍等。这项技术有非常广的用途。

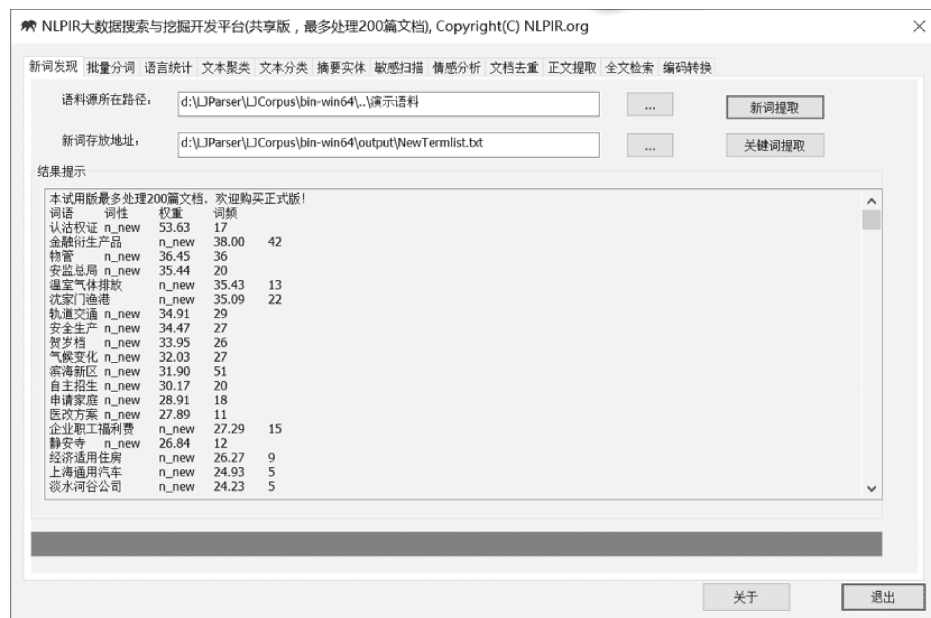


图3 新词发现

这里展示的技术是汉语分词技术。汉语分词是语义理解中最基础性的工作，到目前为止，这项工作已有 17 年的历史，分词效果如图 4 所示，如自动识别一个人的名字、学历和单位，包括英文原型等。这套分词系统已经在全球 40 万个机构使用，华为手机中涉及的语义分析便是使用的这项技术，如短信自动分析时间地点，以及餐饮酒店等。

这里展示的是信息分类过滤技术，系统可将内容类似的文本信息聚合到一起并贴上类别标签。这项技术可以用于文本审查中涉及色情、赌博、毒品等有害信息的过滤，如图 5 所示。

这里展示的是一项基于机器学习的文本分类技术。如图 6 所示，可将类别编成目录文件夹，里面可以放 100 个甚至更多的序列类本，图中展示的是机器自动学习类别特征的过程。

图 7 展示的是经过机器学习后大数据的分类方法，即用深度学习的方法对常规文本进行自动分类，其中交通类文本分类准确度非常高。



图4 批量分词



图5 规则分类



图6 训练分类



图7 分类过滤

不良内容的自动实时智能扫描技术。图8展示的对变形的识别都是音变，语料中并没有直接提关键词，只利用发音扫描到敏感的内容，是语音的智能识别理解技术。其实只要配一个词便可识别各种干扰因素，这样有利于精确打击犯罪，如自动发现赌博、色情等不良信息，寻找需要的信息，挖掘敏感

信息，用户可以通过这种方法得到想要的内容。这项技术的核心特点是智能、速度快，配 100 万个关键词可以做到每秒扫描 20MB 的文本。

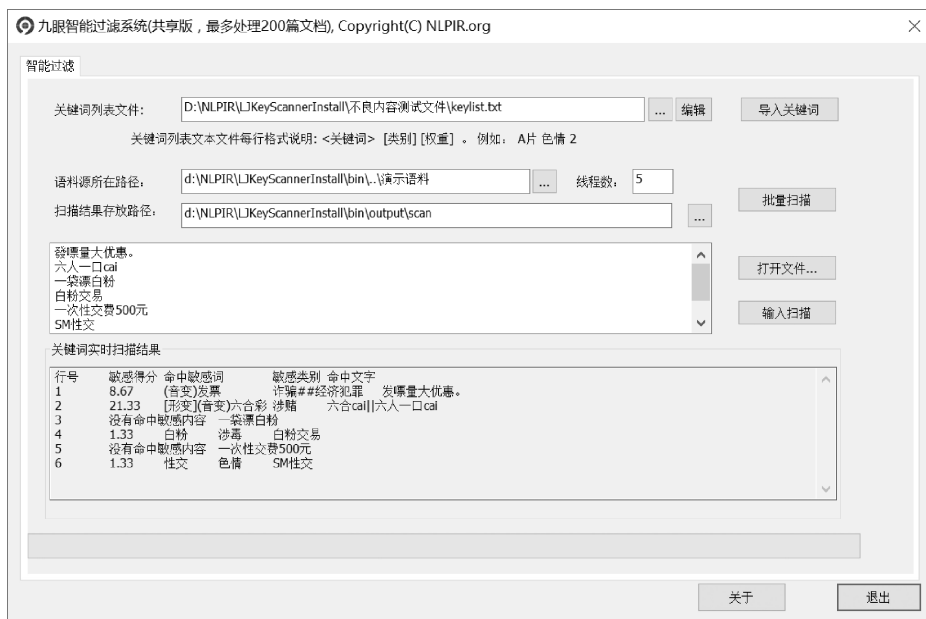


图 8 内容过滤

NLPIR 大数据语义分析技术平台几乎支持所有的开发语言，也支持各种各样的平台，API 可以无缝地融合到客户的各类复杂应用系统之中。

三、大数据精准语义搜索关键技术

JZSearch 语义精准搜索引擎可以采用自然语言的聊天方式，根据语义的知识图谱将某个人的信息展现出来，如图 9 所示，最左边会将某个人相关的十年来所有信息进行聚合运算。

图 10 所示为语义统计分析，是一个时光机技术，我们可以实时计算出每一年的活动、主题，刚才的聚合及每一年的主题，很多词汇都是词典中的内容。值得注意的是，大数据挖掘技术可自动发现某个人的数据关联性。具体原因可以在数据中得到答案。

赋能大数据教育

全国高校大数据教育教学经验谈

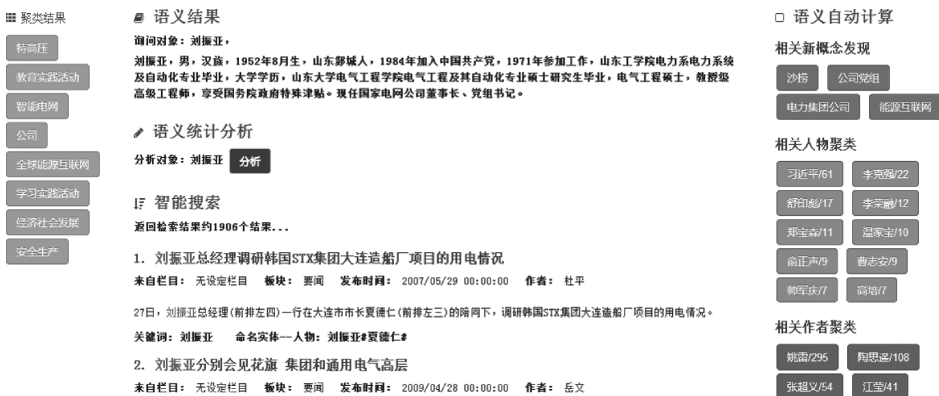


图9 JZSearch 语义精准搜索引擎



图10 语义统计分析

四、大数据语义应用实践

主要介绍以下 4 个案例。

1. 某大厦电力数据挖掘

电力数据基本情况如下：238 个房间，共 300 多天，其间工作日是 256 天，计算其单日用电量。基于这个数据，传统的数据聚合、数据基本分类、数据统计曲线等简单工作在此不再赘述了。

这里涉及的一项关键工作便是计算空置率，空置率的计算对经济预测，尤其是微观经济的洞察和宏观经济的研判具有很强的现实意义。可以看到，这里空置房间的标准是经过大量数据计算出来的。其实在二、三线城市不错的写字楼，其空置率也达到了 32%。除此之外，还可以精确预测每个房间的总体用电情况，由此来推导房间中办公的人数。

2. 95598 客服挖掘

通过对投诉内容的关键词分析，我们可以看到投诉关键词的全国分布、南北方的对比及时段的对比，进而挖掘有价值的信息。

3. 国家电网头条

如图 11 所示，国家电网利用大数据语义分析关键技术打造了一个全媒体个性化智能推荐平台，其中包括全媒体（多位一体、富媒体，如图像、文字、音频、视频、直播等）与云应用（构建了一个开源平台，所有用户、编辑、审核、管理员及技术间的衔接均采用 SaaS 服务）。值得一提的是个性化推荐的尝试和探索（因时因地因人而变），具体而言，指的是不同的人在不同的地方看到的内容是不同的，这里应用了个性化建模、个性化推荐与群体推荐的方法。

4. 公安某局的案件

图 12 展示了一年来盗窃案的总体刻画，其中包括很多有价值的数据。具体以串并案的处理为例，如盗窃三轮车的案件，我们根据案件描述自动从过去的几百万个案件中推荐出前 10 个案件。虽然进行了脱敏处理，但这种处理并不影响数据挖掘。这项工作对于安全的公安部门很有价值。



图 11 电网头条



图 12 公安案件刻画

226

真正有危害的是还不为公众所认知的诈骗案件，值得注意的是有目的地进行诈骗的手法。这种技术适合于对海量数据进行聚合，辅助我们进行综合研判。

通过对同一类案件的人物、地点进行聚合，可以构建一个犯罪地图。犯罪地图分为两种，一种是指犯罪发生地点的地图，另一种是犯罪嫌疑人籍贯地图，通过犯罪地点与犯罪嫌疑人的刻画可以帮助我们发现重大线索。

作者简介

张华平：北京理工大学副教授，博士，研究生导师，知名汉语分词系统 ICTCLAS 创始人，北京市海量语言信息处理与云计算工程中心大数据搜索与挖掘实验室主任，中国中文信息学会社交媒体处理专业委员会副秘书长，北京市顺义区政府专家咨询委员会委员，同时担任辽宁师范大学客座教授，首都师范大学兼职副教授；中国计算机学会青年科技论坛 YOCSEF 委员，中国计算机学会普及工委委员，国家自然科学基金函评专家，北京市重点产业知识产权联盟专家，同时担任《计算机学报》、《计算机研究与发展》、中国科技论文在线等杂志的特邀评审专家。先后获得了钱伟长中文信息处理科学技术奖一等奖，新疆维吾尔自治区科技进步奖二等奖。

油田大数据应用探索

东北石油大学计算机与信息技术学院院长 李春生

一、大数据的基本概念

（一）大数据的定义

大数据是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

针对油田而言，它的数据相对其他行业是较为完整的。油田从勘探开始，一直到采油的整个生产过程中，其所有的数据都有记录，但不同阶段记录的数据形式是不同的。例如，在早期是报表记录，后来通过数据采集形成大粒度的数据，目前通过互联网的应用已形成详细的数据记录。国际上为有效管理目前的油田数据，五大石油公司及一些院校成立了相关组织，用于制定油田数据的模型和标准。目前国内试图引入该模型，但历经十多年后还是停滞了。

中国的油田目前的数据管理基本为两个阶段：第一个阶段是 20 世纪 80 年代，制定了勘探开发数据库，将国内各个行业的工程技术人员、专家等聚集在一起制定中国石油标准。勘探开发数据库是面向主题的数据，根据业务的需要情况组织数据。第二个阶段是目前的石油行业开始了一种新的数据组织方法，对数据进行分层次、多极化管理。

（二）大数据的系统研究维度

大数据应用从以下 3 个方面进行研究：理论方面、技术方面和实践方面。

理论方面：主要是认知必要的途径，即被广泛认同传播的基准路线，如特征定义、价值探讨、目前和未来的发展趋势、大数据隐私等。

技术方面：主要是云计算、分布式处理平台、存储技术、感知技术等。

实践方面：包括互联网的大数据、政府的大数据、企业的大数据及个人的大数据。

二、油田大数据的应用过程

油田大数据的应用过程主要分为大数据来源、浅表特征呈现、数据挖掘分析及其多元化应用研究 4 个方面。

（1）大数据来源：办公数据备案、生产数据积累、数据集成与整合、业务数据处理等来形成原始数据。

（2）浅表特征呈现主要是在原始数据的基础上进行简单的应用和展示，如专项数据成图、模拟仿真建模、列表式展示等。

（3）数据挖掘分析主要是在原始数据积累的基础上进行潜在模式的挖掘、异常追踪预警、知识发现、趋势分析等。

（4）多元化应用研究主要是结合 GIS 应用、数模分析、立体仿真成像、一体化研究等使数据发挥更大的潜在价值。

三、油田数据现状分析

（一）基本概况

（1）涉及领域广。如生产数据、地质地貌数据、措施数据、作业数据、辅助支撑数据等，基本涵盖了油田勘探、开发、生产、装备等所有相关领域。

（2）数据存储形式。包括结构化（关系型数据库）、半结构化（电子表格）及非结构化数据。

（3）逻辑组成。目前国内的油田采取 A1、A2、A3……数据库，按不同专业分出不同的类别，组织油田数据。从数据组织可以区分勘探开发数据逻辑、井下作业数据库逻辑等，过程逻辑结构与面向对象逻辑结构交叉应用。

（二）数据特点

数据完整性较好、数据基数大、数据间业务关系极强、数据准确率较高、

数据更新速度快、数据呈现多样性等。

四、油田大数据的应用方向

油田大数据的应用方向主要有以下 4 个方面：油田大数据集成与处理；追踪式业务及日常工作管理；油田开发生产领域的知识发现及推理应用；与地理信息系统（GIS）、3D 建模技术等结合，以便于辅助决策。

（一）油田大数据集成与处理

油田大数据集成与处理的目标主要满足油田应用业务的需要，油田业务数据结构涉及业务数据模型、采用实用的技术集成和处理过程。由于历史原因，油田所有的信息系统一开始不是集成在一起来做，而是分散孤立的系统，数据类型、表示方式均不相同，如何使所有的油田数据统一化，是集成所做的工作，将分散在各地的数据如何集中起来也是其工作之一。包含的过程有三点：第一，基于油田业务的数据结构、构建、应用领域的业务数据模型；第二，设计数据集成、数据处理过程；第三，完成数据模型的管理及对数据过程的控制。数据集成的意义是强化后期深入数据计算和分析能力，摒弃不相关的数据，提高数据处理速度。这项技术的难点主要是数据处理过程的设计，相对困难一些，针对不同的数据分别一一处理，没有统一的方法。

首先进行模型设计，模型中又涉及两个实体，一个是动态数据实体（反映油田的生产单元），另一个是静态数据实体（描述油田生产单元生产的基本属性）。在两个实体描述过程中需要 3 个数据，分别为开发历程数据、动态管理数据、产量预测及其预警数据。模型设计成功后，面对的问题是如何采用数据集成技术集成相关数据。采用的方法是软件 Agent 技术，在集成的地方开发一个控制中心，生产各个 Agent，根据需要将其分布在数据源所在位置。其中主要包括控制中心（派出 Agent、生产 Agent）、行为业务响应（响应 Agent）、孤立业务处理模型（定时式 Agent）等。

（二）追踪式业务及日常工作管理

追踪式业务及日常工作管理主要采用层级划分的思想，结合实际业务、

工作及其模式管理进行场景式追踪，为决策者提供自上而下的管理支持。包含的过程有三点：第一，底层数据模型的建立，数据的抽取与处理；第二，涉及领域的日常工作及业务管理功能的设计；第三，管理结构与功能对接及其一体化追踪模式设计。这项功能的意义是强化决策者管理能力，提高日常工作效率和管理水平。这项工作的难点是数据模型的建立、标准业务管理及其日常工作的涵盖度。

（三）油田开发生产领域的知识发现及推理应用

知识发现的目标是以知识工程为主导，采用数据挖掘、模式挖掘等技术，实现对油田开发、生产、勘探等领域的专家知识发现及推理应用。包含的过程有数据获取、知识挖掘及知识推理。这项功能的意义是深层发现潜在知识，强化预警及动态分析能力。这项工作的难点是理论方法及应用推广。

（四）与地理信息系统（GIS）、3D 建模技术相结合

与地理信息系统、3D 建模技术等结合，使挖掘出的知识有更清晰的呈现过程，该应用可以使知识更加通俗、易懂。

五、大数据应用案例

（一）聚驱生产动态预警系统

如今通过收集聚驱所有数据，生成相关报表，针对已知的案例挖掘聚驱生产数据呈现的分布模式，进而形成知识。在此基础上，把挖掘出的模式用于对实时生产数据的趋势匹配，如果与趋势吻合，便可以进行预警，通报有关人员采取必要措施。这一系统以数据挖掘为基础，实现聚驱异常生产模式发现，实现单井未来生产状况分析。

（二）天然气安全隐患跟踪系统

天然气安全隐患跟踪系统以天然气安全隐患管理日常业务功能为基础，开发辅助决策、隐患综合跟踪系统，依据自顶向下的管理模式面向决策者开展隐患问题综合跟踪。将基层业务单位对应的隐患问题进行汇总，以业务单

位的模式实现隐患追踪。隐患数据的来源有两种途径：一种是通过人工上报；另一种是通过生产过程实时数据挖掘。决策者可根据跟踪系统追踪到隐患问题的源头。其应用方式是从根节点出发，上位节点统计下位节单位的所有风险隐患。追踪时从本节点开始，进行下级单位追踪，并一直追踪到基层单位的风险隐患情况。基层单位可以进行风险隐患汇总，查看各类风险隐患的分析报告与结果。这个项目通过数据的综合应用达到了科学化管理的目的。

作者简介

李春生：教授、博士生导师，现任东北石油大学计算机与信息技术学院院长，计算机应用技术省级领军人才梯队带头人，全国石油高校信息化协作研究会副理事长，中国教育信息化学会董事，黑龙江省远程教育学会网络教育技术专业委员会副主任委员，大庆市专家委员会委员，国际杂志 *Journal Engineering Letters* 编辑委员会副编辑，《通信学报》审稿专家。1996年获黑龙江省优秀教师，2007年荣获大庆市优秀专家，2009年获东北石油大学教学名师。

广播电视个性化节目推荐系统

中国传媒大学信息工程学院副教授 王 鑫

一、广播电视大数据的由来

传统广播电视收视率调查采用抽样调查的方法，其中包括日记卡和测量仪两种，日记卡数据采集的方式为对 4 岁以上的人员，人工填写，每周进行回收；以 15 分钟为记录单位；数据提供的速度是 15 个工作日，人工进行采集。因此，对于一张记录卡，一人一周的数据采用基于回忆的方法进行数据统计。测量仪则不同，它是采用遥控器特殊操作、仪器调查，以 1 秒为测量单位，以 24 小时为一个统计周期，凌晨固定时间回传，但遥控器家庭成员键配合度低不能实时对数据进行采集。

根据统计学理论，样本数据要达到 1067 个以上，允许的误差才能达到 3% 以下。另外，广播电视对于测量的样本有一定的要求，需要家庭常驻半年以上，周居住超过 5 天的，且家里有电视、经常收看电视节目的人群。

（一）传统收视率调查方法存在的问题

（1）抽样误差。以央视一索福瑞（CSM）为例：全国 650 个城市，总样本户小于 5 万户，平均每个城市不足百户，只有北京、上海、广州等少数城市的样本户达到 500 个。

（2）样本户污染。样本“污染”难以避免。样本户相对固定，隐蔽性差；且日记法、测量仪要求受众参与性强，可以被“收买”。

（3）代表性。样本更换及跟踪难度大，要求人员固定，只能是常驻特定居民。

（4）数据单一。受制于采集手段，往往只有直播数据，缺乏常见的时移、回看。

(5) 只支持传统指标。

(6) 时效性差：遇到特殊情况，需要人工修改数据。

(7) 样本户维护成本越来越高。

(二) 电视大数据采集的要求

1. 用户行为数据

谁在什么时间看了什么频道、节目、页面。

2. 用户特征信息

用户是什么人（地区、年龄、性别、职业、学历、收入）。

3. 媒体资源信息

什么渠道、在什么时间播了什么类型的节目。

4. 用户消费信息

谁购买了什么服务。

5. 服务端业务信息

谁在什么时间使用了什么服务。

6. 终端采集

覆盖三网、多屏；全网数据采集。

(三) 电视大数据分析的要求

不仅需要常规的直播数据，还需包括点播、时移回看、广告业务及其他增值业务等数据。

1. 直播

收视时长、收视率、到达率、接触度、市场份额、观众忠诚度等。

创新指标：节目相对吸引力、收视率分布等。

2. 点播

VOD 业务使用及 VOD 节目指标。

各时段在线户数、在线率、到达户数、到达率、点播户数、点播率、收看时长、页面点击等。

创新的竞争力指标：时间转化率、点击转化率。

按栏目、按供应商分别分析。

3. 时移、回看

业务各时段在线户数、在线率、页面点击率等。

各频道及节目的收视时长、收视率、到达户数、到达率、市场份额等。

4. 广告业务

按各广告位、广告包进行指标分析。

各广告位曝光频次、 $n+$ 曝光率、 $n+$ 到达户数、有效曝光率、有效到达户数等。

5. 其他增值业务

业务各时段在线户数、在线率、页面点击率等。

其他定制指标分析。

（四）广播电视大数据的特点

1. 数据准确、分析计算误差小；公正、抗污染、不易造假

全网海量用户收视数据分析，全方位、无死角，尤其对弱势频道、非黄金时段节目的数据分析更准确，使其数据有意义，更能反映实际情况。

全网双向用户可达千万名以上，用户污染影响微乎其微。

2. 指标更有价值

忠诚度、竞争力指标等，对低收视率的频道和节目，能提供更多参考依据。

3. 能提供舆情分析

根据全网用户的收视行为，结合节目播出信息，可以分析舆情。

涉及舆情及国家信息安全，不建议外资公司参与。

4. 能了解每一个用户的偏好，提供个性化服务

5. 技术难度大

采用传统手段采集海量用户收视数据，成本太高。

采用终端数据回传，需要掌握嵌入式设备、计算机、网络相关的关键技术。

全网收视数据是海量数据（以歌华为例，420 万名用户每天大于 2 亿条），传统的数据库技术无法支撑，需掌握大数据处理系统的架构、算法、专业工具等核心技术。

二、广播电视大数据决策知识系统架构

（一）系统邮件部署的方式

系统硬件部署采取分级的方式，包括数据采集系统、数据传输存储系统及数据分析挖掘系统。

第一级进行数据采集，通过双向网络采集双向机顶盒数据汇集至边际站点。

第二级进行数据传输、存储。汇总边际站点的收视数据、用户特征信息、节目信息及分类数据，形成收视数据库，同时汇总 VOD、时移等业务数据、BOSS 等经营数据。

第三级进行数据挖掘分析。将汇总的各类数据回传至数据分析中心进行数据挖掘，将得到的分析结果以 PC、iPad、手机等终端形式呈现。

（二）广播电视大数据决策知识系统的体系结构

广播电视大数据决策知识系统包含不同的体系结构。

终端层采集数据源包括标清机顶盒、高清机顶盒和智能电视。

数据采集层采集数据包括直播数据采集、点播数据采集、回看数据采集、时移数据采集、广告数据采集、卡拉 OK 数据采集等。

数据存储层通过管理控制层和数据服务层对数据进行综合综合传输和存储。

数据分析曾是软件系统架构的核心部分，包含实时数据分析和非实时数据分析两大部分。

（三）广播电视大数据采集技术

广播电视大数据采集技术采用了 Hadoop 的部署方案，采集服务器将终端机顶盒采集得到的数据回传至中心服务器，并交由不同的服务器分别实现实时分析和 Web 展现等功能。

（四）广播电视大数据存储技术

在大数据存储计算方面，充分发挥了 Hadoop 集群的优势，采用 MapReduce 的分布式计算系统。

（五）广播电视大数据分析挖掘技术

广播电视大数据分析挖掘技术中，采用了 SaaS、R、Python、SPSS 等不同的工具，建立的模型包含支持向量机、决策树、贝叶斯、神经网络等多种不同的算法。

（六）广播电视大数据分析的常规案例

1. 节目基因标签标注

打破了传统的广播电视节目分类体系及“知识树”的结构，采用了扁平化的平行关系，通过从互联网采集节目的标签数据，加上广播节目的标签信息，采用扁平化的标签对节目进行标注。

2. 用户肖像刻画

基于节目标签，定义用户兴趣度；基于节目类型，分析单个用户对哪类节目最感兴趣。

3. 用户分群技术

将用户分群，描述为无收视、低偏好、中偏好、高偏好几类；分析群体偏好，精确至下面的小类。

三、广播电视个性化节目推荐系统

高度信息化的社会每天都会产生海量信息，如何从海量信息中找到用户

所喜爱的节目，为用户进行个性化服务常常困扰着用户。目前互联网各大视频网站纷纷推出个性化节目推荐系统，但广播电视领域还处于一片空白。为此，基于大数据提供广播电视个性化节目推荐系统，为用户提供个性化服务。

在陕西省网中收集到 40 多万个双向用户，每天凌晨将前一天的用户的收视数据上传到北京的机房。在北京，北京歌华有线电视网络股份有限公司目前采集到 420 多万个双向用户的数据。

以某一节目收视板块为例，可以看到不同用户在不同收视日期内看了不同的节目，而不同的节目具有不同的节目特征，在右边的用户偏好板块中，将用户的节目类型进行标签化，可看到用户所喜爱的不同节目类别。当定义到某个用户时，可以看到该用户的节目偏好情况。针对用户的不同偏好提供不同的个性化服务。

本套系统初期选取陕西省网 2 万个家庭用户作为本项目的试点用户，对其免费提供个性化节目推荐服务。未来，将对本套系统进行进一步扩展，扩大用户规模，向陕西 40 万个用户和北京 420 万个用户全面推送本系统，拥有广泛的发展前景。

个性化节目推荐系统面向广播电视各类人群，在为各类人群提供不同服务时，可产生各类回报。对广告商而言，在为用户提供个性化推荐的同时，可精准定位用户偏好，进行广告精准投放。广告收益是本项目的主要收入，节目制作商关心何种节目受欢迎，电视台关心频道收视率如何，从他们那里收取的信息服务费是本项目的增值收入。网络运营商关心 VOD 价值如何，以及有无潜在的增值业务，个性化节目推荐系统可以完全满足用户的需求，视频点播收入是本项目的另一创收点。另外，本项目弥补了广播电视领域的技术空白，政府职能部门予以一定的支持。

中国传媒大学理工学部，旗下成立了大数据分析挖掘研究院。作为其中的负责人，承担了很多课题，包括广播电视个性化节目推荐系统、广播电视舆情分析系统，以及未来将进行的电影影视大数据分析系统。

作者简介

王鑫：中国传媒大学信息工程学院副教授，中传影视大数据研究院技术

负责人，主要研究方向是大数据和信息可视化，先后负责与参与《有线电视用户大数据采集、分析、挖掘和决策支持系统》《基于大数据面向新媒体的节目综合评价系统架构和方法研究》《电影大数据分析决策系统与商业智能》等多个项目，多次获得“校科学技术奖”，并于 2015 年荣获中国广播电影电视社会组织联合会颁发的“2015 年度广播影视科技创新奖”。

大数据与健康

济宁医学院医学信息工程学院院长 孔繁之

一、我国人口亚健康状况

目前，我国人口亚健康状况不容乐观。各国亚健康状况的形成主要是由包括心理、社会、环境、个人生活习惯、行为、气象等各方面的特定内容相互关联所引起的。

如今，社会组织对健康的概念有了一个新的定义：健康不仅仅是没有病和不虚弱，还包括在身体、心理、社会各方面完好的一个状态。

二、“十三五”规划关于推动“健康中国”的五大战略

随着国家“十三五”规划将大数据作为基础性战略资源，全面实施促进大数据发展行动，国家在管理健康方面抓住了大数据的机遇，并结合自身优势不断完善标准规划，以大数据作为信息纽带，利用大数据创造共享价值。大数据展现了传统健康产业对大数据技术的需求，健康产业的变革发生了变化，更重要的是推动了这个行业的发展与成长，促进了新的管理模式的革新。

总体目标如下：以提高人民健康水平为核心，以体制机制改革创新为动力，从广泛的健康影响因素入手，以普及健康生活、优化健康服务、完善健康保障、建设健康环境、发展健康产业为重点，把健康融入所有政策，全方位、全周期保障人民健康，大幅提高健康水平，显著改善健康公平。

三、全生命周期健康管理

全生命周期健康管理，指的是人的生老病死，从出生到婴儿期、幼年期，

再到老年期等整个过程中人的状况管理，包括疾病管理、疾病控制等各种医疗服务。

四、大健康理念

所谓大健康，是指要全社会人群达到健、寿、智、乐、美、德的“六字人生最佳境界”。

大数据与健康的管理理念，主要侧重于未病先防、既病防变，大数据根据中医的意境通过多种方式对亚健康群体实施一系列的健康管理对策，实施以预防为主的干预手段，促使亚健康状态向健康状态转化。

大健康关注的几个主要问题如下：

（一）数据的共享与分析

经过几年的建设，一些区域卫生平台积累了大量的数据，如果能进行有效的共享与分析，将能体现出数据的价值。

（二）大量增长数据的高效存储

接入人口数量增长，采集信息丰富（从基本信息、健康信息到诊疗信息、医学影像等，数据量从几十 TB 到 PB 级）。

（三）众多机构之间快捷可靠传递

各种机构，各种协议，各种业务，如何满足快捷可靠的传递？

（四）安全：服务公众需要的数据和系统安全稳定

信息平台承载了大量居民医疗卫生等相关的关键数据，数据安全的威胁来自物理环境、网络、计算服务、应用等各层面，如何防御？

（五）运营管理：多系统、多业务、多品牌系统

交换机、路由器、服务器、防火墙、虚拟化平台……华为、思科、IBM、VMware……

五、大健康管理流程

大数据支撑下的大健康管理流程，主要侧重于预防、保健、治疗、康复、养老五位一体的大健康管理方案，首先是从身体体检、专科诊疗开始的。

六、大数据带来的挑战

大数据带来的挑战主要来自以下几个方面：

- (1) 医疗资源分布不均衡，基本都集中在发达地区和大城市。
- (2) 医患信息不对称，看病处于无序状态，医患沟通困难。
- (3) 社会老龄化加剧，健康医疗负担加重，急需方便快捷的信息化医疗。
- (4) 信息碎片化，即便在一个地方同级别的医院，患者的诊疗信息也无法共享，更何况全国。
- (5) 医院内部信息化很好，单打独斗的局面继续，没有统一、规范的平台。
- (6) 健康医疗中数据复杂、专业信息带来的专业性壁垒。

七、大健康项目目前的挑战

大健康项目目前的挑战主要与医疗卫生数据的特点相关，具体包括以下六个方面：

(一) 异构性

医疗数据类型的多样化（包括数值型数据、类别型数据、图像、文字、信号、语音、视频），加大了知识发现的难度，使开发基于医疗数据库的通用软件系统较为复杂。

(二) 海量性

医疗工作自身的特点，如病情观察的不可间断、各种医疗检查结果纷繁复杂及存有大量的医学文献专著等。高科技的医学检查设备（如 SPEC、MRI、

PET 等)，每天都会产生数千兆字节数据。

（三）数据特征不显著

数据特征不显著。医疗数据混合了文字、图形等非数值型数据，使得数据挖掘人员并不能很好地找到可以同时反映数据间联系的模型。

（四）难以发掘知识

主观性试验和诊断会带有主观性，难以发掘知识。同一个领域的顶尖专家都会对对方的诊断带有异议，这就会难以整合。

（五）标准化危机

在医学界，很多概念都没有规范。例如，一个简单的概念“结肠腺癌，转移到肝”，就有很多的表达形式；再如，有的中药有很多别名。

（六）伦理性、社会性、法律性

数据归属权问题、数据安全问题、法律诉讼问题等。

八、如何进行大健康管理

大健康管理是指“采用预防—治疗—修养”预防并举的措施，对个人及人群的健康危险因素进行全面管理的过程。通过健康体检、健康评估、健康教育、健康促进、健康保险对个人或群体的健康危险因素进行检测、分析、评估、预测、预防，使公众保持能量平衡、有效运动、量化饮食、身体活动。

数据需要打通，也需要新的共享环节实现数据融合，在整个系统中平滑移动，是大数据针对于大健康的要求。

九、“大健康”服务平台

目前，做一个“大健康”服务平台，主要是针对智慧医疗、老年健康方面，将地方医院、社区、个人的数据进行收集置于信息平台，在信息平台上建立各种与“大健康”所相关的系统，通过健康服务平台，可将医院、社区、

检验中心、养老机构、地方性单位的数据传到服务平台，最终实现对健康人进行预防，对亚健康人进行治疗、调养。在此服务平台上，可进行基本医疗公共卫生、健康随访、风险评估、分疾治疗等业务。初步试想在大健康服务平台上实现健康体检、健康相互、健康服务、健康监测等功能。

“大健康”服务平台已有的系统有以下四个方面。

（一）公共卫生服务系统

搭建公共健康咨询服务平台，提供给公众便捷的健康咨询服务。如健康档案查询、健康咨询服务、自助服务、患者关系管理（患者满意度、纠结与投诉处理、随访跟踪等）。

（二）移动健康

搭建公共健康咨询服务网站，方便公众通过手机即可享受健康咨询服务。如预约挂号、健康档案查询、健康咨询服务、自助服务、患者关系管理（患者满意度、纠结与投诉处理、随访跟踪等）。

（三）移动诊室系统

基于无线访问技术，改变现有医生工作方式，为医院提供全新的医生工作站服务，成为系统亮点之一，医患可以随时互动，随时为患者诊治。

（四）视频探视系统

搭建了一套对核心病房远程探视的服务系统，让患者的家属在病房外的任何地方，都可以对病人实现探视，彻底解决了对 ICU 和 CCU 等核心病房病人的探视矛盾。

十、总结

未来 5 年内，大数据所创造的价值将会达到千亿美元，正所谓得数据者得天下。虽然目前大数据技术在健康大规模应用的条件还没有完全成熟，但随着高速网络、云计算中心等基础设施的日趋完善和大数据技术的不断发展，健康发展的趋势必将是以大数据技术驱动的个性化、创新化、便利化医疗。

作者简介

孔繁之：济宁医学院医学信息工程学院院长，教授，硕士研究生导师。主要研究方向是网络技术与信息安全，健康大数据等。1999 年 10 月作为访问学者由国家公派赴日本筑波大学研修教育技术学。主要讲授《计算机网络》《计算机文化基础》《电路分析》《电工学》《普通物理学》《医用物理学》《Computer Concepts》。多次被评为学校优秀教师、先进科研工作者、先进工作者，荣获济宁市“五一劳动奖章”，指导的学生项目获得首届中国“互联网+”大学生创新创业大赛总决赛银奖。

大数据+旅游——由旅游供给侧需求导出

浙江旅游职业学院教授 任 鸣

一、大数据与旅游的基本概念

个人认为，旅游是人们的一种生活方式，而生活方式是一个动态，它所涉及的数据非常丰富、非常宽泛。提及大数据在旅游中的应用，或旅游需要大数据的支撑，这两个命题都很有意义。事实上，由旅游本身业态所提供的数据是很有限的，且范围比较窄，因此，将旅游业态的数据称为大数据是不恰当的，也是对大数据不正确的一种理解。今天我们讨论的大数据+旅游，实际上是基于“旅游+”和“+旅游”的融合性与关联性大旅游的概念。

二、从旅游宏观供给侧的需求出发谈大数据的应用或需求

所谓旅游宏观供给侧的需求，往往是指旅游的管理能力（社会的管理能力、规划的管理能力）、旅游的服务水平（公共服务水平）、旅游的政策等，也包括旅游资源配置与支配、旅游环境的整治，更谈及有关部门协调，特别是市场监管、市场体系的建立。这些方面都是旅游宏观供给侧的反映，而宏观供给侧恰恰可以提供大量的数据，让这些供给侧更好地为旅游服务、为旅游业的发展扩大需求。因此，我们在浙江省的旅游大数据建设过程中，更多地将这种宏观供给侧的数据作为大数据建设的基础与核心，这也为旅游的需要和发展及旅游建设的进行奠定了基础。

浙江省信息中心从2016年开始着手在这方面进行相应的大数据收集、大数据整理、大数据建设的规范化工作，宏观供给侧大数据的整理工作，主要涉及的是数据的标准化建设及数据的精确化建设。经过实验发现，在旅游宏

观供给侧上有大量的数据需求，同时在提供这些数据的方方面面也有大量的数据存在，这为旅游宏观供给侧的数据化或大数据应用化提供了真正的支持性数据。例如，使用的新的政策、新的样板、新的尝试等，以及发改委审批的和公共服务所提供的各种建设方案、建设思路、建设规划等，都为旅游宏观供给侧提供了许多可参考、可应用、可实施的大数据。

浙江省全域旅游建设，以及小城镇建设的资源配置、环境、市场监管等方面的宏观供给侧需求更加明显，随着浙江省风情小镇、A村景区的创建，往往需要很多部门的支持和协同，而这些部门的支持和协调往往可以通过现有的创新理念增加这方面的信息量和数据量，为风情小镇和A村景区建设提供了强有力的数据佐证和决策。因此，旅游宏观供给侧的需求通过大数据的支撑是很有必要的。

三、从旅游本身的业态来看大数据的重要性

从旅游宏观供给侧方面的需求看到了大数据的重要性，再从旅游本身的业态来看，旅游的六要素：食、住、行、游、购、娱，以及最新定义的六要素等，这些要素本身都是社会的元素，或者都是旅游这种生活方式中必不可少的元素，而这些元素的发展和提供都离不开大数据对这些业态的支撑和支持。在信息化集聚和互联网蓬勃发展、云计算和大数据应用无限广泛的今天，通过对旅游业态中的供给侧需求的整理，为大数据的介入提供了很好的平台。因此，很多传统旅游业态都纷纷介入大数据的洪流之中，不仅为大数据本身提供其自身拥有的数据，更多的是通过大数据来提高其自身品牌和形象。

大数据的应用不仅来自传统的景区、旅行社、饭店等，更广泛地应用在融合性旅游业或旅游+产业，如乡村旅游、工业旅游、休闲旅游等，这些业态更能显现出大数据为旅游产业带来的成果。应该说，旅游供给侧对于大数据的呼唤，以及大数据可以带来的收获是显而易见的，从业态改良到提升都有巨大的影响力。

四、典型案例

上述提到的旅游宏观供给侧和旅游本身出发的大数据，不仅是作为旅游

研究或探讨，更多是作为一个实实在在的旅游供给侧宏观或微观方面的应用。这里以一个典型的案例来讲，从 2012 年开始，开元开创了一种新的酒店业态——曼居酒店，目前已成为曼居酒店集团公司。从第一家绍兴的曼居开始，经过不到 4 年的建设过程，如今已发展了 30 家曼居酒店，它们各有特色，已拥有了 8000 多张床位，拥有了长期的和中层的消费群体，这 30 家酒店的入住率达到 70% 以上，每个房间的售卖平均价达到了 350 元以上，有 40 多家尚在建设中，据公司负责人称，曼居的发展始于大数据（有类似酒店服务的需求）、快速成长于大数据（需求的大数据集聚与区域的释放），而最根本的是得益于大数据（服务形象的社会效益和公司实际的经济效益）。

五、浙江省大数据建设的情况

浙江省大数据建设在旅游大数据建设中位于全国前列，无论战略，还是大数据的应用、标准化建设等都走在前列。在温州召开的全国大数据的会议上也展现了自身大数据建设的风采，在浙江省大数据的建设中有以下几个亮点：

第一，从基础着手。做好宏观基础数据、宏观供给侧数据、旅游业态数据的收集和整理，重点在与旅游相关的食、住、行、游、购、娱、保险等业态都做了大量的数据收集和整理。

第二，在大数据建设中突出标准化的建设。主要在数据的格式、数据的采集、数据的整理及数据的分析方面有自己独特的标准化要求，也为数据的有效利用奠定了基础。

第三，在现有基础数据的前提下，及时整理发布。先后向全省乃至全国发布浙江省的第一次、第二次及第三次的大数据，尽管目前还不完善，但在大数据发布后，的确对于行业有很大的作用。首先，很多业态均可利用发布的数据佐证它们的建设和发展理念是否正确、是否合适；其次，为政府提供有益的发展旅游的佐证，对此，很多地方的旅游部门要求提供更全、更泛的大数据，为其更好地发展当地的旅游业提供良好的决策依据。因此，浙江省大数据建设目前已经认识到了大数据所带来的好处，或者业态可支撑的优势。为此，浙江省旅游局在未来几年要大力打造大数据建设、信息中心建设与数

据中心建设，目的就是使大数据与旅游紧密结合，为旅游更好地服务，提供更有力的保障。目前消费侧正由传统模式向 SaaS 模式转变，而 SaaS 模式中最显著的一点便是过去的欲望改成了搜索，过去的购买改成了行动，过去的最后行动改成了分享，SaaS 模式的消费更多是搜索信息以购买自己需要的东西，新消费中最大的特点是倾向于个人向社会、好友的分享。这些也充分凸显出从消费层、需求方角度看，大数据为改变人们的消费形式提供了很好的平台。旅游业的发展需要大量的游客和需求，为大数据改造供给侧的改革目的是满足更广泛的消费者需要。因此，大数据必定会将消费者的艾德马模式向 SaaS 模式的转变提供有力的佐证和帮助，这也是未来的大数据与旅游之间磨合与切入最好的方式。

希望大家共同探讨，为大数据在旅游供给侧方面的应用和支撑提供更好的建议和意见，为浙江省的大数据建设乃至全国的旅游大数据建设出谋划策，让大数据为旅游增光添彩，为旅游更好地发展提供有力保障。

作者简介

任鸣：中国旅游研究院标准化研究基地首席专家，浙江旅游职业学院教授、高级工程师，浙江大学 EMBA 兼职教授，浙江省旅游标准化技术委员会秘书长，浙江省旅游大数据建设专家。近 5 年来，主要主持和参与两部涉旅国家标准、4 部旅游行业标准和 6 部地方标准编制与修订。参与国家旅游局和浙江省“十三五”旅游标准化规划，完成《中国旅游标准化发展报告》和《旅游标准学》等专著，并且在旅游规划、旅游安全和旅游统计方面有长期的实践与研究。

金融大数据分析与应用

中国人民大学信息学院副教授 许 伟

一、大数据概述

（一）新的时代，人们从信息的被动接受者变成了主动创造者

随着“互联网+”和大数据的快速发展，人们从信息的被动接收变为主动创造。互联网刚刚兴起的时候，出现了一些门户网站，这些网站可以发布新闻等信息，人们可以浏览，也可以简单地留言，虽然那时人们也可以进行交互，但很难做到实时互动。现在我们处于大数据时代，人们使用微博、微信等社交媒体与搜索引擎工具越来越多，人们也就从被动地接收信息转变为主动地发送信息，形成用户之间的高频互动。例如，微信平台本身作为一个社交平台，当人们没有信息传输时，平台仅仅是一个平台，没有人们的相互交流并积累大量的数据，当人们互相发言、讨论，便创造了共同的信息与价值，这些信息与价值也可进行共享、互联互通，更重要的是，在这种情况下，人们变成了信息的主动创造者，人们的行为、痕迹也在网络上进行了留存。

（二）大数据来源

人们在日常生活中可以感受到数据扑面而来，也逐渐意识到大数据对专业领域产生了一定的影响，日常生活中常见的大数据可以从以下几个方面产生。

首先，最容易想到的就是社交媒体网站。例如，Facebook、Linkedin、Twitter、微博、微信这些社交媒体平台产生的数据信息，这些平台不仅包含数千万的用户信息，这些用户还可以主动产生大量的数据信息，不仅有数值型数据，还包括大量的文本信息、图片信息等。

接着，人们平时经常使用搜索引擎来查询想要得到的材料，获取有用的

信息。例如，百度、谷歌等搜索引擎，这些引擎可以查找海量的网页信息，并包含了人们每天关注的热点和搜索行为。

最后，人们日常的金融商务活动在平台上越来越多地进行。例如，电子商务平台和互联网金融平台就可以提供这些服务，这些平台每天都有大量的交易信息，包括商品订单、用户评论等，这些信息可以为平台的持续改进提供有效的支持。

（三）大数据特点

从上述数据来源不难发现，和以前的数据特征相比，目前的数据呈现出更多的特征，这些特征包括数据量大、类型多样、流动性快及价值密度低等，构成了大数据的基本特征。

首先，数据量大。目前的数据呈指数型增长，数据的量非常大，逐步呈现从 TB 到 PB 再到 EB 的增长趋势，日常的服务器很难在一定时间内将这些数据很好地进行处理。

其次，类型多样。目前的数据逐渐从结构化向非结构化、半结构化转变，比如，在电商平台中，以前更倾向于对交易订单等结构化数据进行分析，而现在更多地关注用户评论、商品照片等非结构化数据。

再次，流动性快。以前，人们对于数据处理、管理决策的效率要求没那么高，如今随着互联网的快速发展，人们对决策效率提出了更高的要求，对分析预判的数据快速处理需求增加，在数据快速流动和处理的时代，如何对其进行流动分析和应用变得很有挑战。

最后，数据的价值密度低。从商业价值来看，大数据时代，为了更好地进行管理决策，强相关的数据需要大量弱相关数据进行配合才能取得更好的效果，弱相关数据如果处理得好，会增强强相关数据的预测性，否则，将不能很好地完成特定的任务。例如，人们在购物过程中以前会关注某一件商品的购买关联，现在人们更多地关注某个商品的感受、评价，根据自己的需要选择对自己有用的评价进行商品选择。有关研究表明，在人们的决策过程中，除了对一些商品的销量、价格的关注外，商品评论、用户体验更多地为人们所关心。

这 4 个方面为大数据的几个基本特征，前 3 个方面更多的是从技术角度

来看，而最后一个特征更多的是从商业价值角度来考虑的。因此，如何从大数据中提取有用的信息对目前的数据处理、分析等综合能力要求较强。

二、金融大数据

金融数据分析得到了金融机构、投资者及研究人员的广泛重视，从获客分析到投资决策，从资产定价到风险管理，各个方面都经历了从定性分析到定量分析的逐步过渡，人们试图利用金融模型来得到更好的分析效果。例如，对于银行交易欺诈识别问题，以前的研究更多的是利用历史的交易信息构建欺诈特征，进一步形成反欺诈规则，通过规则判别交易是否存在欺诈。已有的研究中，银行交易信息大多是强相关信息，专业人员利用内部数据来构建金融模型解决需要解决的问题。

随着社交媒体、搜索引擎等大交互数据的出现，使得大交易数据和大交互数据有了融合的可能性，从而可以更好地解决大交易数据中的典型问题。然而，大交互数据虽然有价值，但价值密度低，从中提取有用信息就像大海捞针一样。在这种情况下，如何在大交互数据的处理过程中，迅速提出大交互数据中与大交易数据相关的特征就变得非常关键。针对一个有待建模的金融问题，由于大交互数据量很大，利用大数据处理技术实时响应和快速处理交互数据，配合大交易数据来构建典型特征，有效解决金融问题，为金融业务提供有效的支撑。

三、金融大数据的应用案例

（一）大数据信用评级模型

众所周知，以前的信用评级主要利用用户信用历史来评估，例如，央行的征信中心将人们的贷款违约、信用卡违约集中起来变成信用历史，当进行贷款时根据信用历史估算将来的违约概率，通过概率推断是否可将贷款批复。但某些群体的信用历史是不存在的，如大学生，虽然没有过贷款和信用卡使用记录，但不能认为这些群体是没有信用的，如何通过其他的相关数据推断出这些群体的信用价值？这便是利用其他领域的的数据推断金融领域的信用问

题。这种情况下，如何找到相关数据源并通过数据分析来判断其信用就变得非常重要。事实上，目前的研究表明，购物记录、社交媒体等信息对信用判别具有一定的指示作用，通过数据分析和挖掘方法，从购物记录可以看出用户履约能力和消费能力，通过手机号码可以判别是否是用户本人、是否常用等，通过社交媒体信息反映一个人的行为方式，这些信息都可以从一个侧面来构建新型信用评分模型。

（二）大数据量化分析模型

随着互联网大数据越来越受到人们的关注，在金融量化领域，专业投资者也逐步认识到互联网大数据的重要性，并积极开展大数据量化分析的理论研究和应用实践。人们在投资决策时，容易受到外部信息、情感和人们交互行为的影响，以此为切入点，专业人员构建了不同的基于事件驱动、情感驱动和行为驱动的交易策略并付诸实施。在事件驱动的交易策略构建过程中，利用构建的算法实时动态捕捉热点和突发事件，利用这些事件及事件演进构造相应的买入/卖出信号，尽可能地获得更大收益。同样，通过金融论坛、微博等信息，实时捕捉人们对不同股票的态度，以此为依据构建金融交易策略，这些策略可以更好地获得超额收益。进一步，不同信息源的耦合和集成可以搭建更好的交易策略，进一步指导投资决策。

四、结语

随着大数据的兴起，大数据在金融领域的应用越来越广泛。目前，大数据在金融投资、风险管理等方面取得了一些进展，但研究和实践的宽度和深度仍需要进一步加强。未来，人们会将更多的弱相关数据源引入金融市场，用来分析金融问题，可以更好地进行投资决策。进一步考虑，在解决金融市场问题的同时，大数据技术需要进一步加强，例如，大数据平台和区块链技术为金融市场的有序发展提供了平台支撑。

作者简介

许伟：博士，中国人民大学智慧城市研究中心副主任，中国人民大学信

息学院信息系统与大数据应用实验室主任、副教授，博士生导师。中国系统工程学会信息工程专业委员会秘书长，《系统工程学报》编委，*Journal of Systems Science and Information*、《系统工程理论与实践》《管理评论》等国内外期刊客座主编。主要研究领域为金融管理、电子商务、智慧城市、信息系统。主持国家自然科学基金、教育部人文社会科学研究规划基金、北京市自然科学基金、北京市社会科学基金及金融企业合作项目多项，在 *Annals of Operations Research*、*Decision Support Systems*、*European Journal of Operational Research*、*IEEE Trans. Systems, Man and Cybernetics*、*International Journal of Production Economics*、*Production and Operations Management* 等国内外期刊会议上发表研究论文 90 余篇，出版专著 5 部，获得北京市科技新星、北京市优秀人才、北京市哲学社会科学优秀成果奖等多个奖项。

健康大数据的降维问题

吉林大学计算机科学与技术学院教授 周丰丰

一、大数据在健康信息学中的挑战

以深圳市为例，如果每年采集的健康体检数据，包括电子病历数据、心电图数据、基因组和医学影像数据全都保存在一起，就需要很多硬盘来保存，仅仅保存就需要这么多的硬盘和空间，所以需要合理的方法对这些数据进行合理的处理，而不是将所有数据全部拿出来进行分析。

计算挑战性的问题，主要原因在于这些健康数据的体量，一个人会产生数千万维的数据，由于数据采集的成本问题，每个人需要几万到几十万的数据量级，如果完整做下来，需要相应的资金，因此，不太可能采集超过 100 个人以上的数据量级。此时在计算上存在一个“大 p ，小 n ”的挑战，主要原因在于当数据、特征很多，但人群个数很少时，如果将所有的数据全部进行数据建模，便会发生上述问题。具体体现在当一个数据挖掘的模型出来之后，如果全部使用一千万个特征，这样的模型不够稳定，即一个新的人出现后，会发生无法预测的困难，这样的困难在统计学上存在一个需要解决的挑战，建模所需要的特征数比样本数少，才会是一个很好的模型，要想解决这个问题，就要在几千万的特征中挑出几个特征，使得其特征数少于样本数。也就是说，我们需要将特征个数降下来，当特征个数少于样本个数时，这样的数据模型会比较稳定。

我们可以将其中一些无关的基因去掉，从生物角度来看，如果研究对象是肺癌，并不是所有的基因都与肺癌相关，可将无关基因、冗余基因、噪声基因去掉，使得实际使用构建模型的基因个数少于实际样本的个数，避免后面的过滤盒问题，即来了一个新样本，老样本训练构造的模型无法使用的挑战性问题。

从技术角度来看，在整个解空间上，每一个点都是数据的一个解，都是一个特征子集。少数几个特征代表一个点分布在纵坐标上，纵坐标越高，说明其性能也就越好，未来构建模型越准确，这就需要找出其中最好的点。通过每次都找最好的邻居，只能找到局部的最优点，而无法找到全局的最高峰。

这也描述了从计算角度上来看特征选择问题的主要计算挑战，原因在于整个空间上需要找到一个一开始不是最好的点。通过周围不断选择较好邻居的过程，可以找到实际中全局最优的点。工作的假设是一开始找到不是最好的点的过程，在全局过程中可能会找到最好的点。通过一开始找到最好的点的方式实际是过滤法的一个常用策略，这种策略通常结果很好，但效果不是很好。

二、三个臭皮匠赛过诸葛亮

从理论上构造出一个实际的例子，有两个有疾病的人和两个健康的人，假设考察他们的 F1 基因和 F2 基因，发现 F1 基因的数值是 1010，F2 基因的数值是 0110，F1 基因单独做分析时的区分能力是 50%，与随机加分的效果是一样的。F2 也是 50% 的分类准确性，这时只需要构建一个易获的分类模型，两个数值如果不同，则取值为 1，如果相同，则取值为 0，这样便可将其分类准确性达到 100%。因此，两个单独效果差的两个基因组合到一起可达到很好的效果，这就需要考察单个效果一般的特征能否组合在一起进而更好。

如何将特征子集中不需要的特征去掉呢？可以考虑如下方法，假如一个特征与另一个特征是线性相关的，即 $F1=2F2$ ，从数据分析的角度来看，在两个线性相关的特征中去掉一个，保留另一个，在整个模型构建中是不影响区分能力的，也就意味着这两个特征之间是冗余的，去掉其中一个不会产生任何影响。

我们使用的是 Maximal Information Coefficient (MIC) 系数，这个系数主要描述的是不同数据相关性的相关系数能否合理描述出来，如果两个特征形成线性相关时，所有的相关系数都可以很好地描述出来，如果是平方、指数的复杂性系数，其描述能力差别非常大，但 MIC 系数可将其很好地描述出来。因此，用 MIC 系数来描述两个数据之间的相关程度。

基于 MIC 相关系数的特征选择算法，通过这种算法将很多特征个数降下来，同时保证分类的效果很好，其中重点描述的是算法。

特征选择算法主要分为过滤法、包装法及嵌入法。嵌入法是指特征选择过程是嵌入在一个不断优化分类准确性的过程，最常用的是过滤法和包装法两大类特征选择算法。通过采集的数据集，详细考察通过 MIC 的特征选择算法能否从至少 4000 多个特征到数万量级降到特征个数差不多的量级上，并保证很好的分类效果。通过一个两步优化的算法，将数万量级的特征降到了与样本特征个数差不多的量级上，最后也很好地保证了预测的准确率。

第一步优化的主要策略，目的是将可能的冗余性去掉。首先考察的是信息的相关性，考察与类标相关性足够大且大于一个预值的特征保留下来，主要目的是考察与类标信息不相关的，考察一个特征是否具有不错的效果，并保留下来。信息冗余实际上是 F_i 和 F_j 两个特征，如果 F_j 特征的类标足够相似，且相似程度大于 F_i 与类标间的相关性，进而 F_i 与 F_j 间的相关性又大于 F_i 与 C 类标的相关性，也就意味着 F_j 的特征已从分类的效果和相关信息足够覆盖 F_i 特征所描述的指标，这时便将 F_i 去掉，用 F_j 代表。因此，通过信息覆盖的原则，将 F_j 特征保留下来，具有足够的信息相关性，通过这两个思路便可将较多的与类标不相关的、信息间的冗余信息去除。

McTwo 算法首先使用第一步的 McOne 算法，选择具有足够分类效果的特征子集，但在此过程会发现特征个数仍然居多，在几十到上百个量级上，接下来使用最优优先搜索的策略，将近邻分类算法嵌入，接着基于平衡准确性的指标 $(S_n + S_p) / 2$ ，通过去一法验证策略作为优化目标，来发现是否为一个最好的特征子集被合理选择出来。整个算法的实验流程是基于交叉验证，与多个过滤法和包装法的特征选择算法进行对比，最后算出综合分类效果，分析哪一个特征子集的分类效果最好。特征子集分类效果的评判指标，真阳性、假阳性、真阴性、假阴性等，可以计算出敏感性、准确性、平衡准确性、相关系数等指标。分析发现，McTwo 算法有必要进行两步，性能也比其他包装法、过滤法好。

McOne 第二步加上与否对于算法性能的比较，准确性方面与 McTwo 的一步算法相似，McOne 算法的准确性与 McTwo 效果相似，但 McTwo 得到远远小于 McOne 算法的特征个数，这也说明用同样的特征可以得到同样的分类准

确性，特征数越少，模型越稳定，同时在未来临床使用时的成本也越少。

与现有包装法的算法，其中包括 CFS、PAM、RRF 做对比时，会发现两个例子数集：胃癌数集与一型糖尿病的数集，McTwo 算法在准确性方面是效果最好的，综和而言与现有算法相差不多，同时实际的特征个数少很多。进而发现，在整个数集上，算法会比大部分数集效果好。McTwo 的算法个数远远少于其他算法的个数，CFS 是最好的，基本是算法的 20~30 倍的特征个数，考虑到下方的指标时，准确性减去特征的个数除以 100，多用一个特征性能，指标会降低一个百分点，综合而言，McTwo 算法的特征个数和准确率都是一个很好的兼顾。

与过滤法做对比时，McTwo 算法与其他算法相比效果相差不大，胃癌数集相对偏好，神经网络方面与算法结合是最好的。在不同的过滤法对比时，算法会得到给定的特征个数，根据给定的特征个数来选择过滤法的特征个数，当然部分数集上也有不如现有算法的情况。

三、外部调查验证效果

从统计学上来看，交叉验证分为内部交叉验证与外部交叉验证，内部交叉验证是指先选择一组特征，然后将整个数集进行交叉验证。而外部交叉验证是将数集随机分成几个等份，接着选择特征做测试，分析其是分成一步来做，还是分成两步来做的过程。在测试的 3 个数集上，算法有很高的一个性能，其方差比较小，即上下波动较小，因此，稳定性比较好。但组合在一起效果很好的一组特征并不一定需要采取过滤法效果很好的特征，其中甚至选择了最后的几个特征。

参考文献

- [1] J Liu, C Xu, W Yang, Y Shu, W Zheng, Fengfeng Zhou. Multiple similarly-well solutions exist for biomedical feature selection and classification problems[J]. Scientific Reports, 7 (1): 12830.
- [2] Y Ye, R Zhang, W Zheng, S Liu, Fengfeng Zhou. RIFS: a randomly restarted incremental feature selection algorithm. Scientific Reports 7.

- [3] R Ge, M Zhou, Y Luo, Q Meng, G Mai, D Ma, G Wang, Fengfeng Zhou. McTwo: a two-step feature selection algorithm based on maximal information coefficient[J]. BMC bioinformatics, 17 (1): 142.

作者简介

周丰丰：教授，博士生导师，中国科学院百人计划，IEEE（美国电气和电子工程师协会）高级会员。周丰丰博士的团队主要从事健康大数据挖掘核心算法的研究。已发表学术论文 54 篇，其中包括 SCI 索引 41 篇，主要作者论文（通信或者第一作者）44 篇。根据 SCI 数据库统计，总引用次数 711 次，他引次数 534 次。

基于认知心理学的大数据可视化设计研究

上海交通大学媒体与设计学院副教授 萧 冰

如何使大数据可视化信息更容易理解？数据可视化会涉及哪些要素？从视觉设计的角度来看，共包括图形、色彩和认知心理三方面的内容。

大数据可视化是通过将复杂的数据转化为可以交互的图形，帮助用户更好地理解分析数据对象，发现并洞察其内在规律。从而可以更方便地进行数据分析，促成合作与信息共享，以及使终端客户具有处理信息的能力。从学科支撑的角度来讲，认知心理学和图形设计是数据可视化的两大基础学科。

一、可视化变量

可视化变量分别为位置、形状、方向、颜色、纹理、明度等级、尺寸。它们的感知性质，从认知心理学角度分析，可以表示数据的联系性、选择性、次序性、数量性。

二、形状变量

就形状变量而言，遵循“少即是多”的原则。在信息可视化图形中，用的最多的是线形、三角形、正方形、圆形等抽象的几何图形。

根据“群魔殿”认知理论，认知的过程分为4个层次，分别为“映像鬼、特征鬼、认知鬼、决策鬼”。

“映像鬼”的工作过程是视网膜记录外界形象的过程。

“特征鬼”是对这个形象进行分析，如英文字母、垂线、水平线、斜线、直角等，每个“小鬼”负责报告一种特征。

“认知鬼”是从“特征鬼”的反映中寻找与自己负责的识别图像相关的特征，并大声喊叫出来，符合的特征越多喊声越大。如R这个字母，有竖线、

圆弧形、直角和斜线等特征，符合其中部分或全部特征的有 D、P、R 三个字母，负责这些字母的“小鬼”都会叫出来，但负责 R 的“小鬼”发现的特征最多，叫喊声就更响亮一些。

“决策鬼”根据“认知鬼”的喊声大小判断需要识别的图形。通常来说，图像本身越简单越容易识别。

三、方向变量

方向与形状密切相关，环状结构没有明确的方向性，便于用户自主发现联系；而像箭头、手臂这样的图形，其方向性是很明确的，设计师要对其中的联系先有一定的预想；动态图形运动的方向则能够凸显出各部分的联系性。

四、色彩变量

相对于图形而言，色彩系统是一套完善的体系。在表达次序性、选择性上的作用突出，色彩包括 3 个维度：色相、明度、纯度。通常所说的赤、橙、黄、绿、青、蓝、紫，这些都是不同的色相，所有的色相都是不同波长的可见光呈现出的颜色，将这些颜色组合在一起构成环形就是一个色相环。而黑色到白色的色阶变化称为明度变化。从最纯的颜色到它同等明度的灰色，它们之间的变化称为纯度的变化。根据色相、明度、纯度 3 种维度描绘出来的体系就是色立体。

不同的色彩在人们的心理上会形成不同的影响，如蓝色会让人想起冰天雪地，会感觉到寒冷，橙色、红色会让人想起火焰，感觉到温暖，称为暖色。此外，还有轻和重、愉快和伤心、前进和后退等不同的色彩心理。不同色彩间的组合可引起色彩识别度之间的差异，如最醒目最容易识别的图形是黄底上的黑色图形，排在第二位的是黑底上的黄色图形。这和发展进化历史密切相关，如老虎的斑纹、蜜蜂的条纹都是黄色和黑色的结合。

色彩在表达次序上很有优势，我们经常会根据色相来判预警的等级，蓝色预警、黄色预警、橙色预警、红色预警。但根据赤、橙、黄、绿、青、蓝、紫的顺序会发现黄色和蓝色之间缺少了绿色，这是由于绿色是在所有可见光波中居中的色彩，在心理上给人的感觉最放松，不会引起警惕的感受，所以，

不适合做预警色彩。

色彩往往比图形更容易挑动观众的视觉。从色彩上讲，除明视度、调性外，还具有诗意。

可视化信息的形态通常具有 4 种类型：图表式、抽象类比式、比喻式、寓意式。

抽象类比式包含了金字塔结构图形、环状结构图形等常见的结构图形形式；比喻式包含的树状结构、根状结构也是我们常用的图形。

MIT 的学者用眼动仪观察用户观看可视化数据的过程，发现以下几点特征：第一，看一眼就能记住的可视化图像中要含有可以被记住的内容，要具有视觉关联和语义关联。第二，标题和文字是可视化图像中的关键要素，帮助人们回忆所看到的内容。第三，象形图不会阻碍人们记忆或理解可视化图像。第四，冗余信息有助于回忆和理解可视化图像。

另外，有意思的一点是，最不可识别的可视化图像 54% 来自政府部门（美国），他们采用的可视化图像往往采用相同的模板和类似的美学特征，因此，容易造成识别的混乱。

作者简介

萧冰：毕业于英国伯明翰艺术与设计学院。上海交通大学媒体与设计学院设计系教师，院长助理。中国设计师协会（香港）会员。曾供职于上海美术设计公司。为 30 多家企事业单位等做过品牌形象整合设计，曾参与世博会中国馆及主题馆主题策划。专业方向：互动新媒体艺术、视觉传达设计。

大数据与生物信息学的应用研究与实践

青岛大学数据科学与软件工程学院教授 李劲华

一、相关背景

(一) 生物信息学产生背景

众所周知，生物信息学是 20 世纪 80 年代末随着人类基因组计划的启动而兴起的一门畸形交叉学科，通过对生物学实验数据的获取、加工、存储、检索与分析，进而达到解释数据所蕴含的生物学意义的目的。当前生物信息学发展的主要推动力来自于分子生物学，生物信息学的研究主要集中于核苷酸和氨基酸序列的存储、分类、检索和分析等方面。因此，目前的生物信息学可以狭义地定义为将计算机科学和数学应用于生物大分子信息的获取、加工、存储、分类、检索与分析，以达到理解这些生物大分子信息的生物学意义的交叉学科，实质是理论概念与实践应用并重的学科。

生物信息学的产生与发展已有 30 多年的时间，美国人类基因组计划中对基因组信息学的定义是一门学科领域，包含基因学组信息的获取、处理、存储、分配、分析和解释的所有方面。自 1990 年美国启动人类基因组计划以来，人与模式生物基因组的测试工作发展极为迅速，提前完成了 40 多种生物的全基因测试与工作。截至目前，仅登录在美国 GeneBank 的 DNA 系列总量便超过 70 亿碱基对。此外，迄今为止，已有一万多种蛋白质的空间结构以不同的分辨率被测定。基于 cDNA 序列测试所建立起来的 EST 数据库已超过数百万条，在这些数据基础上派生、整理出来的数据库已达 5000 多个。

这一切构成了一个生物学数据的海洋。这种科学数据的极速和海量积累在科学发展史上是空前的，但数据并不等于信息和知识，当然，它是信息和知识的源泉，关键在于如何从中对其进行挖掘。与正在以指数方式增长的生物学数据相比，人类相关知识的增长却十分缓慢。一方面是巨量的数据，另

一方面是人们在医学、药物、农业与环境等方面对新知识的渴求，这些新知识将帮助人们改善其生存环境和提高生活质量。这就构成了一个极大的矛盾。这个矛盾就催生了一门新兴的交叉科学，这就是生物信息学。

信息学大数据研究工作主要以分析海量多元组学数据为目标，组学大数据为生命科学带来了前所未有的机遇，在研究基因功能、疾病机理、精准医学等方面具有重要意义。大数据的规模性、多样性、高速性等特征为生物信息学带来了新的挑战，在数据计算方面，亟须解决中小实验室对计算资源的弹性需求；在数据分析方面，亟须多组学整合分析体系解决生物学问题。缺乏相应的生物学工具是大数据时代生命科学领域面临的主要瓶颈。

（二）青岛大学生物信息学研究背景

（1）2009 年，位于武汉大学的国家软件工程重点实验室在青岛举办暑期学校，首次听到西方学者提到计算机以生物学跨学科研究，主要包括基因测序、生物大数据可视化等。

（2）2011 年起，青岛大学与深圳华大基因研究院联合创立青岛大学华大基因创新班，培养大数据时代生物基因组学、生物信息学领域的拔尖创新人才。在大学生入校后一个月的时间内，从全校 9000 多名不同专业的学生中择优挑选 30 人，按照厚基础、宽口径、综合式、国际化的要求，在学科基础课和专业课程阶段设有两个选课模块，一个是医学检验，另一个是信息处理。

（3）2016 年，与青岛大学医学部教授合作，共同申报获批了生物信息学二级学科的硕士点，研究方向主要是序列和基因组学的分析、药物研发、生物学网络整合、数据挖掘和数据分析（主要是在生物学应用领域）、生物信息学软件方法学的研究。

二、生物信息学研究的主要内容、主要问题和关键技术

（一）生物信息学研究的主要内容

1. 基因组学研究

基因组学包含了构成和维持一个生命有机体所必备的基本信息，由细胞内进行的多种分子生物学反应将这些信息转换为真正的生命现象。基因

组的一部分编码蛋白质和 RNA，其他部分调控这些大分子的表达。表达的蛋白质及 RNA 折叠为高度专一的三维结构，在体内的特定位置上实现这些功能，这些过程的大量细节都是在分子生物学的实验室里揭示出来的，形成大量数据，存储于数据库中。生物信息学试图从这些数据中提取新的生物学信息和知识，是一门植根于全面深入的实验事实和数据的理论生物学。

2. 生物信息的收集、存储、管理与提供

包括建立国际基本生物信息库和生物信息传输的国际网络系统；建立生物信息数据质量的评估与检测系统；生物信息的在线服务；生物信息可视化和专家系统。

3. 基因组序列信息的提取和分析

包括基因的发现与鉴定，如利用国际 EST 数据库和各自实验室测定的相应数据，经过大规模并行计算发现新基因和新 SNPs，以及各种功能位点；基因组中非编码区的信息结构分析，提出理论模型，阐明这些区域的重要生物学功能；进行模式生物完整基因组的信息结构分析和比较研究；利用生物信息研究遗传密码起源、基因组结构的演化、基因组空间结构与 DNA 折叠的关系，以及基因组信息与生物进化关系等生物学的重大问题。

4. 生物信息分析的技术与方法研究

包括发展有效的能支持大尺度作图与测序需要的软件、数据库，以及若干数据库工具，如电子网格等远程通信工具；改进现有的理论分析方法，如统计方法、模式识别方法、隐马尔科夫过程方法、神经网络方法、复杂性分析方法、密码学方法、多序列比较方法等；创建一切适用于基因组分析的新方法、新技术。包括引入复杂系统分析技术、信息系统分析技术等。

5. 应用与发展研究

汇集与疾病相关的人类基因信息，发展患者样品序列信息检测技术和基于序列信息选择表达载体、引物的技术，建立与动植物良种繁育相关的数据库，以及与大分子设计和药物设计相关的数据库。

（二）研究问题

1. 生物大数据的存储与管理

包括生物大数据的存储结构、存储标准、管理技术等，生物大数据数量大、结构复杂、存储标准多样，存在非结构化数据、半结构化数据和结构化数据等多种数据结构，如何选择分布式文件系统、分布式数据组合、分布式并行数据库系统也是生物大数据存储与管理技术的主要问题之一。

2. 生物大数据可视化

生物大数据由于数量巨大，具有普遍生物意义，合理的可视化可以帮助生物学家快速理解和分析生物数据。

3. 生物大数据的分析与处理

整合多组学数据进行计算分析以解决实际的生物问题。

（三）关键技术

生物大数据领域中的关键技术有如下几种。

1. 生物大数据标准化和集成、融合技术

研究组学数据、医疗数据和健康数据集成融合关键技术，研究开发组学、医疗和健康数据信息模型与集成引擎，研究基于国内外标准规范的消息、文档等接口实现技术，基于下一代互联网技术网络安全技术和高吞吐量传输技术。

2. 生物大数据表述索引、搜索与存储访问技术

重点突破生物大数据资源描述和并行访问技术，构建生物大数据高效索引和可靠可扩展存储管理系统，基于语义的生物大数据资源检索、生物医疗数据关联搜索等关键技术，建立生物大数据资源搜索与获取服务系统。

3. 心血管疾病和肿瘤疾病大数据处理分析与应用研究

分别针对心血管疾病和肿瘤疾病，集成电子病历、图像影像、临床检验数据等多类型数据（覆盖 50 万以上个体人群，总数据量 50TB），开展医疗大

数据的处理、存储、分析、应用研究，为提高重大疾病的诊治水平提供大数据支撑。

4. 基于区域医疗与健康大数据处理分析与应用研究

选择覆盖 100 万以上个体人群，总数据量不少于 100TB 的区域医疗与健康数据，通过处理、存储、分析、整合，构建面向健康服务的知识库及支撑平台，并提供应用服务。

5. 组学大数据中心和知识库构建与服务技术

集成包括基因组、蛋白质组等组学数据，总数据量不少于 100TB，至少 60% 以上的数据提供对外访问，重点突破个人基因组可视化技术，组学注释与疾病风险评估技术，建立组学大数据知识库及搜索引擎、数据挖掘和可视化分析平台。

作者简介

李劲华：青岛大学数据科学与软件工程学院，计算机科学博士，教授，研究生导师，副院长，青岛大学“软件工程”学科负责人。

大数据治理规则体系构建研究

中国人民大学信息资源管理学院教授 安小米

本报告关于大数据治理规则体系构建研究问题的提出主要来源于前期多个相关研究课题的研究发现。其中，我主持的国家社科重大课题“国家数字档案资源整合与服务机制的研究”，从2014年4月到2017年10月对国家数字资源管理的现状进行了初步调查，涉及104个机构，54家政务资源中心和技术供给机构，50家档案馆，采访了300多人，覆盖了9个省、3个直辖市、1个自治区、18个市和8个区级机构，是了解我国政府信息资源管理现状及其存在问题和原因的基础，为选题研究的必要性和可行性论证提供了重要的支持。此外，之前我主持过的北京市经济和信息化委员会项目《政务信息资源公益性开发和再利用管理研究》、国家发展改革委员会重大研究课题《信息化（大数据）提升政府治理能力》子课题“大数据背景下政府数据资源的可持续管理与利用机制研究”、深圳市经贸信息委项目《深圳市政府数据开放行动计划》和《深圳市促进大数据发展行动计划》、中国人民大学科研项目《基于大数据的智慧城市服务关键技术研究及典型应用》子课题“智慧城市大数据集成技术与信息资源整合方法”和所参与过的中国工程院咨询研究项目中李国杰院士主持的子课题“智能城市信息环境建设与大数据”都是本课题研究的基础。本报告的主要观点主要源自洪学海教授主持的国家自然科学基金项目重点培育项目“面向政府决策的大数据共享与治理机制”中我所负责的子项目“大数据治理规则体系研究”的研究构想。

一、课题提出的研究背景

（一）国际背景

全球范围内开放政府数据，运用大数据完善社会治理，提升公共服务能

力，推动经济发展已经成为一种国际化趋势。例如，美国 2012 年发布《大数据研究和发展倡议》，2014 年发布《大数据：抓住机遇与守护价值》和《大数据与隐私：技术视角》，2015 年发布《美国创新战略》，2016 年发布《信息作为战略资源管理》《大数据问题：是包容还是排除的工具》《联邦大数据研究与开发战略规划》，2017 年发布《推动美国政府数字化战略》等把数据服务能力提升被纳入数字创新国家战略。澳大利亚 2013 年发布《公共服务大数据战略》，2015 发布《大数据优化实践指南》，2016 年发布《开放政府国家行动计划 2016—2018》》，数据治理能力提升被纳入数字转型国家战略。英国 2013 年发布《把握数据带来的机遇：英国数据能力战略》，2016 发布《数字英国战略》，数字治理、数字素养和数字技能培养被纳入数字国家战略。截至 2014 年 2 月 10 日，全球已有 63 个国家加入开放政府合作伙伴组织，2014 年欧盟发布《数据驱动经济战略》，越来越多的国家加入开放数据运动。大数据涉及的多元主体共治难题也越来越多地受到学界的关注。

（二）国内背景

我国政府 2015 年发布了《促进大数据发展行动纲要》，首次将大数据发展列入了国家战略，2015 年还发布了《关于运用大数据加强对市场主体服务和监管的若干意见》，2016 年发布了《关于促进和规范健康医疗大数据应用发展的指导意见》和《大数据产业发展规划（2016—2020 年）》。从行动计划的落地来看，全国各地纷纷都出台了大数据发展行动计划，根据不完全统计，有 20 个省 25 个地市已发布了地方大数据发展行动计划，13 个省市成立了 21 个大数据管理机构，8 个国家大数据综合实验区和 15 个大数据交易所纷纷建立。然而，国家大数据发展政策的制定本身尚待加强多元主体的合作，省级及地市大数据行动计划实施方案的制定本身亟待加强大数据与信息资源管理协同创新的运行机制构建。

（三）相关文献述评

第一，从大数据研究现状来看，关于大数据的认识存在多种视角和不同动因，归纳起来，包括新思维、新方式、新能力、新权益、新资源和新基础设施 6 种视角，如表 1 所示。

表 1 大数据的代表性观点

视角	代表性观点	代表性观点出处
新思维	数据驱动、数据管理、数据决策、数据创新文化；创造新价值的源泉、创造新价值过程、改变市场、组织结构，以及政府和市民关系的方法	舍恩伯格，2013； 邬贺铨，2014； 马建堂，2015
新方式	大数据将带来一场管理革命，无论是企业界、学术界，还是政策界，都将受到重大影响；大数据的复杂型特征需要采用经济高效和创新方式处理，也将会把我们的洞察力、决策力、执行力和竞争力提高到新的高度；对国家而言，大数据意味着国家治理体系的重构、变革与升级一场管理革命，产生新管理模式；对企业而言，大数据将掀起一场管理革命，产生新商业模式	安德鲁和麦卡菲，2013； 马建堂，2015； 鲍尔，2015
新能力	具有融合（Fusion）、云计算（Cloud），洞察力（Insight）与预见性（Foresight）四大能力；数据驱动的生产能力；建设数字国家需要的数字连接力、数字技能、数字基础设施、数字业务、数字治理、数字经济、在线生活和工作的网络空间安全和安全环境维护等能力	倪光南，2016； 《英国 2015—2018 年数字经济战略》，2015； 《数字英国战略》，2017
新权益	大数据的数据权属影响国家数字主权，国家信息领空安全、影响数字空间国家竞争力；影响公民知情权、隐私权、记忆留存权或遗忘处置权等个人权益；如果没有制度、法律、文化的支撑，大数据技术本身是对人文、人类生存、社会伦理和正义及民主的巨大威胁	李国杰、程学旗，2012； 马建堂，2015； 黄欣荣，2015； 《大数据问题：是包容还是排除的工具》，2016； 于文轩，2017
新资源	大数据是国家战略资源；是可再生和可循环利用资源，可以衍生新产业和新应用；信息作为国家有价值的资源和政府的战略资产，应该可获得、可发现和可使用，具有互操作性和开放性	怀进鹏，2013； 邬贺铨，2014； 杨善林、周开乐，2015； 《开放数据政策：信息作为资产管理》，2015； 《信息作为战略资源管理》，2016
新基础设施	大数据资源建设作为其信息化基础设施和智慧城市基础设施建设的有机组成，包括完善政府电子政务系统、数据开放平台和大数据应用平台建设、采用基于大数据的创新方法解决社会问题	《美国创新战略》，2015； 《数字英国战略》，2017

第二，从大数据治理体系构建的研究现状来看，也存在不同的认识，从宏观层认识，认为大数据治理体系构建应该是一种概念体系构建，包含目标、

权力、对象、问题等（郑大庆，黄丽华，张成洪，张绍华，2017；郑大庆，范颖捷，潘蓉，蔡会明，2017）；有人认为它应该是一种体系框架，包括战略方针、组织架构、责任分工等（张绍华，潘蓉，宗宇伟，2016）；也有从中观层认识，认为它应该包含管理机制（胡志伟，汪振强，2014；潘永花，2015）、数据治理计划（Soares，2014；程广明，2016）、数据全面管理的部署（Mohanapriya et al，2015）；还有从微观层次认识，认为大数据治理体系构建应该是一种治理策略、治理程序（Malik，2013），对数据全生命期经济有效的管理（Loshin，2013；Tallon，2013）；有人认为是为大数据治理行为提供有关的技术工具运用（Loshin，2013；Tallon，2013；梁芷铭，2015）等不同的观点。大数据治理研究目前存在学科比较独立、学科间互动较少、研究维度较单一、缺少学科视角互联、研究层面各自为政，以及较少覆盖宏观、中观和微观多层级合的研究局限。

第三，我国的大数据治理研究与国外多学科合作多元化综合集成的研究视角相比，还存在以技术为主导，缺少多样化的治理复杂性问题思考，对个人信息安全保护与合法合规合理利用考虑不足的局限。大数据治理研究中如何平衡个人隐私保护与与个人信息合法合理利用提高公共服务供给力、如何平衡大数据精准治理与负面影响风险管控等问题亟待研究。国外大数据治理研究关注的焦点有面向大数据主体多元化的整合要求倡导协同治理和深度融合（Williamson，2014；Jordan，2015；Liao，2015），面向大数据资源供给客体的功能拓展，以公共服务为核心逐步向各领域扩散（Bhimani & Willcocks，2014；White，2014；Dhotre，Vayena，2015；Williamson，2015），基于大数据利用流程的分析与实践，多样化的治理方式，在管理方式、法律规范、大数据策略方面均有突破（Bonilla，2013；Clark，2013；Davenport，2014；Kemp，2014；Lemieux，2014）；基于组织机构业务活动的资源配置与支撑要素，安全保障程度不断加深，实现从个人用户隐私到企业治理风险控制的范围细化，并与个人伦理道德规范、企业文化等方面逐渐融合（James，2014；Williamson，2015）；政府数据治理的应用案例研究，覆盖社会治理领域和公共服务领域（Clark，2013；White，2014）。我国涉及大数据治理的研究关注的议题如下：对大数据的表示方法治理（李国杰，程学旗，2012；黎林峰，陈宝权，2015）；对大数据信息的有效融合（李国杰，程学旗，2012）；对大

数据高效率低成本的存储治理（李国杰，程学旗，2012；怀进鹏，2013）；对非结构化和半结构化数据的高效治理（李国杰，程学旗，2012）；大数据治理活动，用城市大数据构建新的计算理论与方法，提升城市智能化与数据开放化治理水平（怀进鹏，2013；王宇德，2014；黎林峰&陈宝权，2015）；大数据治理与知识服务（刘澜冰，2015；吴若溪，2015；左欣蕊，马众，2016）；大数据治理服务机理（李国杰，程学旗，2012；张勇进，2015；张春景，曹磊，等，2015；于施洋，王建冬，等，2016）；大数据驱动政府治理转型研究（安小米，2015；陈之常，王淼，2015；汪玉凯，2016）。

二、大数据治理面临的挑战与大数据治理规则体系缺失问题的提出

从我国大数据治理的实践现状来看，在大数据发展应用背景下，我国政府数据资源的管理面临数据客体、管理主体、管理活动、管理风险变化 4 个方面的挑战，大数据治理规则体系缺失问题日益突出，亟待加强研究（安小米，2015；安小米，毛春阳，2015）。

（一）大数据治理的客体

从管理的客体来看，数据资源面临形态上从过去的线下转变为线上线下的融合，从单一到多样，从静态到动态，从结构化为主转向非结构化为主等转变；从价值来看，从单一转向多元，面向政府、面向企业、面向公众的各种价值多样化的需求，从信息传递转向信息增值再用；从战略地位来看，从过去面向组织层转到面向行业层、地域层、国家层、国际层、“一带一路”成为国际竞争性资源；从数据权属关系来看，从简单转向复杂具有不确定性特征，数据的所有权、处置权、利用的许可权及隐私保护权等法律依据问题亟待研究。

（二）大数据治理的主体

大数据发展应用背景下，数据主体涉及的利益相关方越来越多，政府从传统的数据权利所有者、控制者和监管者逐步转向数据权力的协调者和社会

协同治理的服务者，从部门利益转向了政府的整体利益，以及智慧城市和信息惠民社会利益的最大化。从过去的信息孤岛转向跨层级、跨领域、跨地域、跨系统、跨部门和跨业务的信息资源融合与创新服务，多利益相关方的合作日益重要，相关的复合型人才的培养越来越迫切。

（三）大数据治理的活动

从大数据治理涉及的数据资源管理活动来看，采集、存储、利用、维护方面也发生了变化。采集从单一来源转向多源异构，从基于目标的局部采集转向基于场景的全面采集，从行业转向地域和国家统一大数据资源体系建设；存储从分布式冷备份存储转向热备份和云存储，从可信数字仓储建设转向可信区块链平台建设；利用从机构内部共享，逐渐扩展转向跨地域、跨领域、跨层级、跨系统、跨部门和跨业务共享，不仅要强调互联互通，更要强调互信互任互动，解决零距离、零材料、零跑腿的问题；在数据的维护上，亟待从实现一次性跑腿或零跑腿的服务需求出发，建立数据全生命期、全流程、全要素的综合集成管理机制，构建数据链生态维护、互联网+社会协同治理的创新服务模式。

（四）大数据治理的风险

在“一号一窗一网”面向老百姓的信息惠民服务中，数据的汇聚与政府外包服务和 PPP 融资方式也带来了数据利用和再利用中的数据权力、权利和权益失控风险，个人数据安全保护和分级分类合法合理利用的规则亟待制定。数据整合、大数据分析结果的公开都可能带来对数据所有权、数据利用权方面的影响，特别是碎片化个人信息再次整合后，对个人身份的再定位可能对个人隐私暴露带来的风险迫切需要研究。

（五）大数据治理规则体系构建研究的必要性

当前我国大数据治理中存在跨领域、跨地域、跨层级、跨系统、跨部门、跨业务的数据链断裂问题，亟待建立覆盖数据全生命期的数据资源供给服务体系，有效规划和融合政府数据采集、信息公开、数据开放、大数据应用等数据资源管理活动。大数据治理中存在数据资源管理规则体系和服务规则体

系缺少统筹规划的问题，当前共享机制、安全机制、开放机制各自为政，相互隔离的问题亟待解决，面向信息惠民服务，亟待构建互联互通互信互认互动的融合机制，贯穿数据全生命期，如《中华人民共和国国家安全法》《中华人民共和国网络安全法》《网络产品和服务安全审查办法》《互联网新闻信息服务管理规定》《政务信息资源共享管理暂行办法》《关于推进公共信息资源开放的若干意见》等，亟待加强多利益相关方的统筹协调，以及跨领域和跨部门的合作实施。大数据资源谁来建，亟待解决建设主体领导力与协同能力问题；大数据资源从何来，亟待解决共享与开放规则和依据规范问题；大数据资源如何用，亟待解决隐私与安全风险管控制度与保障措施问题；大数据资源如何可持续再用，亟待解决处置与留存合法合规合约问题。

此外，我国大数据资源的可持续再用保障体系亟待建立，在完成国家发改委的项目中发现，数据资源可持续再用目前的法律依据存在严重的冲突问题，从数据可追溯、可关联可管控的需求来看，国家相关政策文件尚缺少将信息作为资产的战略管理策略，文件中涉及最多的信息活动是共享、公开与安全，缺少数据资源数字连续性管理的战略意识，缺少覆盖数据资源全生命期的全程性管理规划，缺少对数据资源长期保存及可持续再用的管理准则，缺少数据资源资产化管理制度和风险管理规范，缺少数据资源跨领域、跨地域、跨层级、跨系统、跨部门、跨业务共享的协同管理和创新服务信息化发展战略。

从数字社会和数字经济的发展需求来看，电子文件（包括电子证据、电子证照、数字凭证等原生数字文件）的合法性和可信性亟待解决，当前《中华人民共和国电子签名法》（2004）和《关于办理刑事案件收集提取和审查判断电子数据若干问题的规定》（2016）难以支撑政府数字转型及可信数字政府建设，大量被访机构采用双套制和双轨制管理电子文件，数字原生的电子文件被打印为纸质保存，再电子扫描提供利用，资源浪费问题严重，电子文件合法性问题严重影响电子取证和数字身份认同，影响跨层级、跨地域、跨系统、跨部门、跨业务的电子证照互认及其无纸化公共服务。

从数字信息资源的可用和可再用性来看，《中华人民共和国档案法》（2003）、《中华人民共和国政府信息公开条例》（2007）、《中华人民共和国保守国家秘密法》（2010）、《政务信息资源共享管理暂行办法》（2016）存在信

息主体权利和责任划分缺少互联和互认的合作问题，实际部门难以适从，如按谁的信息谁公开，档案部门保存的政府历史档案信息的开放存在依据性文件冲突问题，缺少可操作性法律依据，导致综合档案馆馆藏档案开放率普遍显著下降的负面影响问题。如何构建大数据发展应用背景下数据资源多元主体合作共治的协同创新规则体系，构建数据资源多元主体合作共生的管控准则；构建数字资源利用和再利用多元主体合作共赢的许可机制和共享契约合同规范亟待研究。大数据治理中存在 3 个关键性问题，即面向社会治理的协同创新能力提升，跨部门的数据连续性管理计划制度规则亟待制定；面向社会治理的公共服务能力提升，跨系统的智能化开放数据算法设计规则亟待研究；面向社会治理安全能力提升，跨业务的自动化个人信息保护和安全监管规范操作规则亟待建立（安小米，2015）。

三、大数据治理规则体系构建的研究构想

从推动大数据与实体经济深度融合的十九大国家战略要求出发，从建设全国一体化的国家大数据中心，推进技术融合、业务融合、数据融合，实现跨层级、跨地域、跨系统、跨部门、跨业务的协同管理和服务的国家战略需求出发，洪学海教授主持的国家自然科学基金项目重点培育项目“面向政府决策的大数据共享与治理机制”中我所主持的子项目“大数据治理规则体系研究”课题提出了以下 3 个科学问题，拟克服现有大数据治理相关研究的局限，针对大数据活动中治理规则体系缺失及其研究不足的问题：谁来负责大数据治理体系的构建？什么是具有可持续发展特征的大数据治理生态体系？如何构建具有可持续发展特征的大数据治理生态体系？子课题提出了以下 3 个研究目标拟回答三个研究问题：“（Who）”多元主体的协同创新共同体联盟机制构建研究；“（What）”跨部门、跨行业、跨领域协同创新的生态治理体系构建研究；“（How）”互联、互通、互信、互认的多利益相关方协同创新知识服务平台构建研究。课题提出要从两个维度考虑问题的解决：一是大数据治理的治理对象的复杂性问题，即大数据数据资源的体量复杂性、价值复杂性、管理复杂性和计算复杂性问题；二是大数据治理的治理活动及其服务的复杂性问题，即多主体、多需求、多过程、多模式的治理活动及多样化服务方式。

大数据治理规则体系构建应该包括以下两个方面：一个是构建大数据价值多维、主体多元、过程多样的长效治理规则体系；二是构建从数据孤岛到“互联网+”数据连续性管理和知识服务联动的规则体系。在构建长效治理的规则体系中，应该包括跨学科大数据治理多维度连接的共识规则、跨领域大数据治理多元主体联盟的共生规则、跨部门大数据治理多层级联通的共享规则。

（一）大数据治理规则体系构建的研究问题

这个子课题存在以下 3 个难点问题待研究：

“Who”（谁来负责），涉及大数据生态治理体系的社会建构共治规则及其协同治理体系构建问题，以期优化政府大数据资源价值实现的路径、资源配置方式及其数字治理能力提升。

“What”（数据价值合法合规合理实现的依赖性要素是什么），涉及风险评估的共识准则和安全管理体系统构建问题，以期提升各种要素互联互通中的安全保护和合理利用及其数字管控能力。

“How”（如何处理多利益相关方的多种多样的利用目的和需求），需要连接主体、客体、活动和风险管控的技术支撑，即协同创新体的共识、共生、共享和共赢准则的技术实现及架构，以期提升多利益相关方的数字素养、数字技能和数字服务能力。

（二）大数据治理规则体系构建的研究思路

本子课题以目前大数据治理过程中出现的数据本身、管理主体、管理过程的变化所带来的大数据挑战为对象，通过实践调研及问题梳理，分析大数据本身的属性特征、管理主体及管理过程的转型变化特点；以协同创新理论、公共价值理论、数字连续性理论和多元论为依据，分别从大数据供给治理体系、大数据资源服务治理体系、大数据资源保障治理体系 3 个维度构建政府大数据治理规则体系生态环境，如图 1 所示。

这个子课题的创新点在于以公共价值论指导服务治理体系构建，旨在解决主体联盟的主体角色转型问题；以数字连续性理论指导资源保障治理体系构建，旨在解决过程联通的管理过程多样性问题，以资源价值多元论

为指导，旨在解决要素连接的供给治理体系中的知识服务方式多样化问题。针对政府大数据治理规则体系构建研究多元主体共治复杂性难题，首次在具体研究过程中通过引入信息技术、信息管理、信息通信技术、社会技术、信息资源管理等多学科融合视角，将这些学科领域按数据—信息—知识—行动的发展历程分层次地融入到政府大数据治理规则体系构建的研究中，旨在为大数据治理规则体系构建提供一个综合集成的、自适应的、知识服务导向的生态环境，为数据驱动的公共服务能力提升与社会治理现代化能力提升提供一种便利性服务和安全性保障的数字治理方式，为数字经济和数字社会可持续发展提供一种社会架构与技术架构互联互通、互信互认的知识服务能力，如图 2 所示。

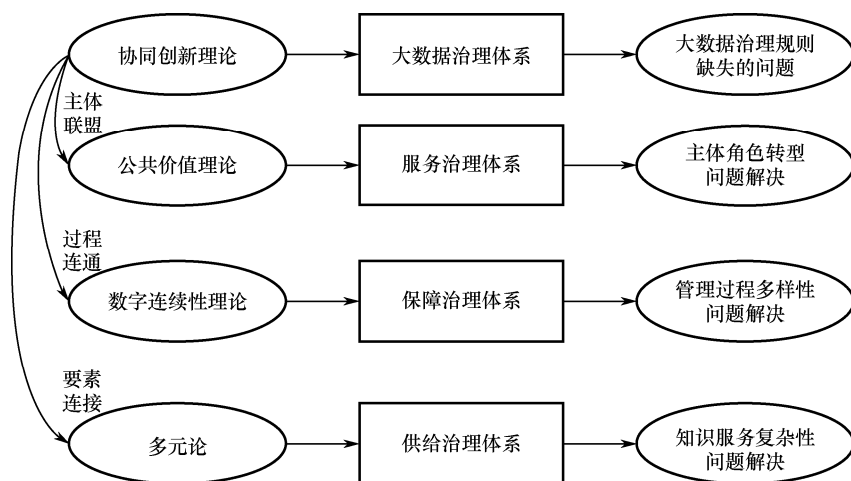


图 1 大数据治理规则体系构建的跨领域融合研究思路

（三）大数据治理规则体系构建的合作研究倡议

大数据治理规则体系构建亟待跨学科、跨领域和跨部门的合作研究，亟待大数据与信息资源的多利益相关方合作协同，构建多元主体共认的数据互通准则，构建多维度数据共享的数据互联契约，构建多样化数据共生的互认规范。大数据治理规则体系构建研究亟待建立综合集成的、自适应的、知识服务导向的生态环境，基于政、产、学、研合作的大数据和信息资源协同创新共同体构建可建议为一种有效的生态路径。

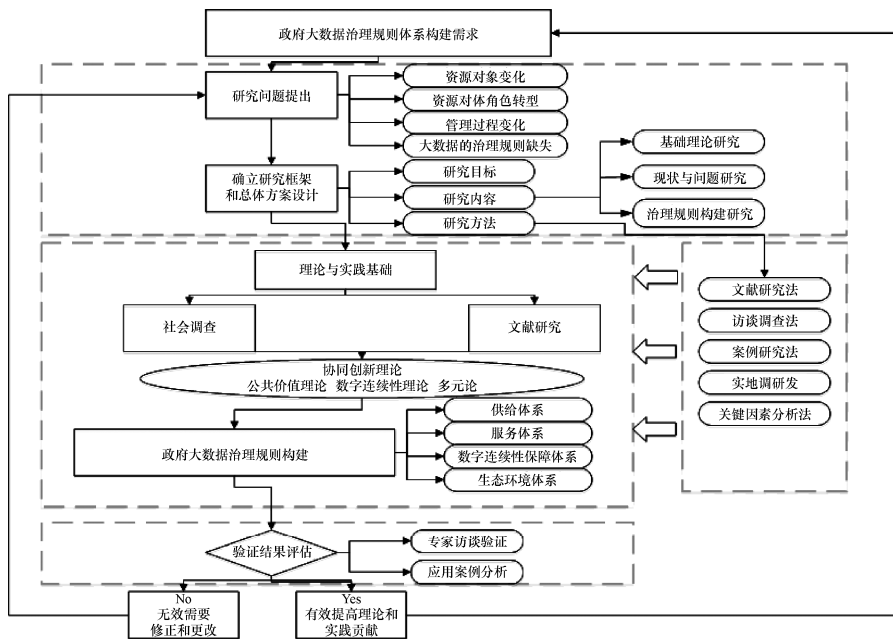


图2 大数据治理规则体系构建的研究路线

2016 年举办的中国信息资源管理论坛暨新型智慧城市与大数据资源管理研讨会上，由中国人民大学信息资源管理学院发起的中国大数据信息资源协同创新共同体成立，从 5 个角度提出了开展大数据治理规则体系合作研究的倡议：①推动大数据与信息资源的互联互通、互信互任互动规则构建；②完善协同治理和服务标准规范的制修订工作机制；③倡导基于数据全生命期、全过程、全面质量管理的数字连续管理和知识服务联动机制构建；④强调多利益相关方共同维护数据主体的权力、权利和权益，为社会治理创新服务做出积极贡献；⑤大力推动大数据发展应用，共创大数据应用新思维、新方法、新模式、新技术、新产品和新服务，提升大数据技术创新与应用服务和产业服务发展互动的敏捷性。

借此机会，感谢微信平台给予我机会倡导大数据与信息资源协调创新共同体构建，跨学科、跨领域和跨部门合作开展大数据治理规则体系构建研究，欢迎多利益相关方参与并支持我们的研究，多方对话，共同应对大数据发展应用带来的变化，共同孵化我国的大数据治理的规则体系，共同维护我国大

数据治理规则体系构建的生态环境。感谢我的博士生董宇、宋懿和张宇杰为申请国家自然科学基金重点培养项目子项目付出的努力。

参考文献

- [1] 安德鲁, 麦卡菲. 大数据: 一场管理革命[J]. 射频世界, 2013 (2): 34-36.
- [2] 安小米. 现代国家治理的云端思维: 信息治理能力与政府转型的多重挑战[EB/OL]. 人民论坛学术前沿, 2015-02-06.
- [3] 安小米, 毛春阳. 大数据时代的政府信息治理[J]. 中国建设信息, 2015 (12): 58-59.
- [4] 蔡钰. 地方政府大数据治理能力现代化建设研究[J]. 湖北省社会主义学院学报, 2017 (3): 60-63.
- [5] 曾凯. 大数据治理框架体系研究[J]. 信息系统工程, 2016 (11): 130-131.
- [6] 陈之常. 应用大数据推进政府治理能力现代化——以北京市东城区为例[J]. 中国行政管理, 2015 (2): 38-42.
- [7] 程广明. 大数据治理模型与治理成熟度评估研究[J]. 科技与创新, 2016 (9): 6-7.
- [8] 单志广, 房毓菲, 王娜. 大数据治理: 形势、对策与实践[M]. 北京: 科学出版社, 2016.
- [9] 范灵俊, 洪学海, 黄晔, 等. 政府大数据治理的挑战及对策[J]. 大数据, 2016, 2 (3): 27-38.
- [10] 高小平. 从传统治理到大数据治理——阅读《大数据时代的国家治理》[J]. 广州公共管理评论, 2015 (3): 287-296.
- [11] 胡志伟, 汪振强. 关于大数据治理的研究与分析[J]. 时代报告, 2014 (7): 177.
- [12] 怀进鹏. 大数据是国家战略资源[J]. 中国经济和信息化, 2013 (8): 49-50.
- [13] 黄欣荣. 大数据技术的伦理反思[J]. 新疆师范大学学报(哲学社会科学版), 2015, 3: 46-53+2.
- [14] 黎林峰, 陈宝权. 创新大数据理论和方法提升城市智能化水平[J]. 中国建设信息, 2015 (3), 22-25.
- [15] 李璠, 柯丹. 构建大数据能力核心引擎, 主动拥抱金融科技创新——中国光大银行大数据治理体系规划与实施[J]. 中国金融电脑, 2017 (5): 27-30.
- [16] 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 战略与决策研究, 2012, 27 (6): 18-24.

- [17] 梁芷铭. 大数据治理: 国家治理能力现代化的应有之义[J]. 吉首大学学报: 社会科学版, 2015, 36 (2): 34-41.
- [18] 刘澜冰. 大数据时代的知识管理平台构建[J]. 情报工程, 2015 (6), 109-112.
- [19] 马朝辉, 聂瑞华, 谭昊翔, 等. 大数据治理的数据模式与安全[J]. 大数据, 2016, 2 (3): 83-95.
- [20] 马建堂. 大数据: 大变革、大机遇[J]. 中国经贸导刊, 2015 (28): 22-23.
- [21] 马亮. 大数据治理: 地方政府准备好了吗?[J]. 电子政务, 2017 (1): 77-86.
- [22] 潘永花. 领导干部应关注大数据治理的哪些理念[J]. 决策与信息, 2015, 12: 29-30.
- [23] 汪玉凯. 大数据时代政府的治理创新[J]. 信息安全与通信保密, 2016 (6): 44.
- [24] 王宇德. 企业大数据治理研究[J]. 互联网天地, 2014 (1): 20-24.
- [25] 维克托·迈尔·舍恩伯格. 生活、工作与思维的大变革[M]. 周涛, 译. 杭州: 浙江人民出版社, 2013.
- [26] 邬贺铨. 大数据思维[J]. 科学与社会, 2014 (1): 1-13.
- [27] 邬贺铨. 关于大数据的若干思考[J]. 中国信息化, 2014 (9): 3-7.
- [28] 吴若溪. 大数据时代知识管理的新风向[J]. 学理论, 2015 (19): 161-162.
- [29] 杨善林, 周开乐. 大数据中的管理问题: 基于大数据的资源观[J]. 管理科学学报, 2015, 5: 1-8.
- [30] 于施洋, 王建冬, 等. 国内外政务大数据应用发展述评: 方向与问题[J]. 电子政务, 2016 (1): 2-10.
- [31] 于文轩. 大数据之殇: 对人文、伦理和民主的挑战[J]. 电子政务, 2017 (11): 21-29.
- [32] 张春景, 曹磊等. 公共文化服务大数据应用模式与趋势研究[J]. 图书馆杂志, 2015(12): 4-8.
- [33] 张绍华, 潘蓉, 宗宇伟. 大数据治理与服务[M]. 上海: 上海科学技术出版社, 2016.
- [34] 张勇进. 智慧政务与政府治理转型[J]. 传媒, 2015 (5): 21-24.
- [35] 郑大庆, 范颖捷, 潘蓉, 等. 大数据治理的概念与要素探析[J]. 科技管理研究, 2017, 37 (15): 200-205.
- [36] 郑大庆, 黄丽华, 张成洪, 等. 大数据治理的概念及其参考架构[J]. 研究与发展管理, 2017, 29 (4): 65-72.
- [37] 左欣蕊, 马众. 大数据时代下的知识管理方法探究[J]. 中国管理信息化, 2016, 19 (9): 198.

- [38] Bhimani A., & Willcocks L. (2014) . Digitisation, “Big Data” and the transformation of accounting information[J]. Accounting & Business Research, 44 (4): 469-490.
- [39] Bonilla, D. N. (2014) . Information management professionals working for intelligence organizations: ethics and deontology implications[J]. Security & Human Rights, 24 (3-4): 264-279.
- [40] Clark, E. E. (2013) . Reflecting inward and looking outward: future trends impacting corporate governance research and practice[J]. Global Journal of Comparative Law, 2 (2): 115-146.
- [41] Crampton J W. Collect it all: national security, Big Data and governance[J]. Geojournal, 2015, 80 (4): 519-531.
- [42] Davenport, T. H. (2014) . How strategists use “big data” to support internal business decisions, discovery and production[J]. Strategy & Leadership, 42 (4): 45-50.
- [43] Dhotre, P., Shimpi, S., Suryawanshi, P., & Sanghati, M. (2015) . Health Care Analysis Using Hadoop [J]. International Journal of Scientific & Technology Research, 4 (12): 279-281.
- [44] Englmeier K. Role and Importance of Semantic Search in Big Data Governance[M]// Big-Data Analytics and Cloud Computing. Springer International Publishing, 2015.
- [45] Fryman L, Lampshire G, Dan M. Chapter 9 – Governing Big Data and Analytics[J]. Data & Analytics Playbook, 2017: 233-242.
- [46] James, R. (2014) . Out of the box: big data needs the information profession-the importance of validation[J]. Business Information Review, 31 (2): 118-121.
- [47] Jordan, S. R. (2014) . Beneficence and the expert bureaucracy[J]. Public Integrity, 16 (4): 375-394.
- [48] Kam. B. Options for Smart City Development[R]. Renmin University of China, January, 2015.
- [49] Kemp, R. (2014) . Legal aspects of managing big data[J]. Computer Law & Security Report, 30 (5): 482-491.
- [50] Lemieux, V. L., Gormly, B., & Rowledge, L. (2014) . Meeting big data challenges with visual analytics: the role of records management[J]. Records Management Journal, 24 (2): 122-141 (20) .

- [51] Liao, Z., Yin, Q., Huang, Y., & Sheng, L. (2015) . Management and application of mobile big data[J]. International Journal of Embedded Systems, 7 (1): 63-70.
- [52] Loshin D. Chapter 5-Data Governance for Big Data Analytics: Considerations for Data Policies and Processes[M]//Big Data Analytics. Elsevier Inc. 2013: 39-48.
- [53] Malik P. Governing Big Data: Principles and practices[J]. IBM Journal of Research & Development, 2013, 57 (3/4) .
- [54] MERTILOS, A. Development of a Capability Maturity Model for Big Data Governance: Evaluation in the Belgian Financial Sector [M/OL]. [2017-10-09], <http://scriptiebank.be/en/node/4912>.
- [55] Mohanapriya C, Bharathi K M, Aravinth S S, et al. A Trusted Data Governance Model for Big Data Analytics[J]. International Journal for Innovative Research in Science and Technology, 2015, 1 (7): 307-309.
- [56] Power, Daniel J. Decision Support Use Cases with High Volume, High Velocity, and High Variety Data [R]. Istanbul, 2015: 21-22.
- [57] Priebe T, Markus S. Business information modeling: A methodology for data-intensive projects, data science and big data governance[C]//IEEE International Conference on Big Data. IEEE, 2015: 2056-2065.
- [58] Soares, S. Big Data Governance: An Emerging Imperative. California[M], United State: Mc Press, 2014.
- [59] Tallon P P. Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost[J]. Computer, 2013, 46 (6): 32-38.
- [60] White, S. E. (2014) . A review of big data in healthcare: challenges and opportunities[J]. Open Access Bioinformatics, 2014 (6): 13-18.
- [61] Williamson, B. (2014) . Knowing public services: cross-sector intermediaries and algorithmic governance in public sector reform[J]. Public Policy & Administration, 29 (4): 292-312.
- [62] Williamson, B. (2015) . Governing software: networks, databases and algorithmic power in the digital governance of public education[J]. Learning Media & Technology, 40 (1): 83-105.
- [63] UK: Digital economy strategy 2015-2018 [EB/OL]. (2015-02-16)[2017-11-05]https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/404743/Digital_Economy_

Strategy_2015-18_Web_Final2.pdf.

- [64] UK: Digital Strategy [EB/OL] (2017-03-01) [2017-11-05]<https://www.gov.uk/government/publications/uk-digital-strategy/uk-digital-strategy>.
- [65] US: Open Data Policy—Managing Information as an Asset[EB/OL] (2015-05-26) [2017-11-05]<https://project-open-data.cio.gov/policy-memo/>.
- [66] US: A STRATEGY FOR AMERICAN INNOVATION.[EB/OL] (2015-10) [2017-11-05]
https://obamawhitehouse.archives.gov/sites/default/files/strategy_for_american_innovation_october_2015.pdf.
- [67] US: Big Data: A tool for inclusion or exclusion? [EB/OL] (2016-01) [2017-11-05]
<https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf>.
- [68] US:Managing Information as a Strategic Resource, [EB/OL] (2016-7-28) [2017-11-05]
<https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/OMB/circulars/a130/a130revised.pdf>.

作者简介

安小米：英国利物浦大学哲学博士。中国人民大学信息资源管理学院教授，博士生导师。数据工程与知识工程教育部重点实验室（中国人民大学）信息资源管理研究团队负责人、中国人民大学电子文件管理研究中心国际动态研究室主任、中国人民大学信息资源管理学院智慧城市研究中心数据治理研究室主任。

社会网络中顶点相似性度量方法研究与应用

燕山大学信息科学与工程学院教授 郭景峰

一、什么是社会网络

社会网络是以社会生活中人类或组织的行为活动为研究对象，将其抽象为相互作用的个体组成的网络，这些网络一般用图的方式表达。社会网络研究是理解社会现象、预测人类行为、分析社会结构的重要工具。

二、社会网络研究现状

1. 社会网络分类

现有社会网络通常包括以下几类。

(1) 传统社会网络：通常以单一类型的顶点和单一类型的边来表示实体之间的联系，即传统的图论中所涉及的内容。

(2) 符号社会网络：同时具有单一类型的顶点对立关系的边，可以通过顶点之间的联系表达出朋友或敌人、喜欢或讨厌等对立关系，一般将友好关系用正边来表示，敌对关系用负边来表示，这是目前研究的新领域。

(3) 多模社会网络（异构社会网络）：一个网络中包含不同类型的实体（顶点）和不同类型的实体（顶点）间关系。

2. 社会网络的相关研究

社会网络的主要研究内容包括链接预测、社区发现、社区演化、信息传播、影响最大化等。

3. 研究基础

上述研究大多以顶点间的相似性度量方法为基础。

三、基础理论——集对分析理论

我们使用的主要方法为集对理论，集对理论是我国学者赵克勤教授提出来的，其主要观点是将所研究的对象看成确定、不确定的关系。

1. 集对分析理论——核心思想

任何事物都是由确定性和不确定性构成的，不确定性在一定条件下可以转换成确定性。

2. 集对

由一定联系的两个集合构成的对子，表示为 $H=(A, B)$ ，表示两个集合之间有一定的联系。

3. 联系度

重点定义两个研究对象的属性集分别为 A 和 B ，两个集合的同异反关系用图 1 表示。两个集合 A 和 B 中，公共的部分称为它们具有相同属性，其中一些完全对立的称为相反属性，其余部分称为相异属性。这里的“异”在一定条件下有可能转换成“同”，也有可能转换成“反”。

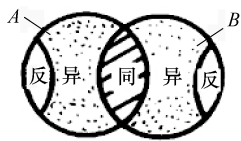


图 1 A 和 B 的同异反关系

两个研究对象的联系度形式上表示为

$$u = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j$$

式中， S 表示相同的属性个数， F 表示相异的属性个数， P 表示相反的属性个数， N 表示所有集合中元素的总数。

4. 基于集对的社会网络模型实例

可见，集对理论及联系度可以应用于社会网络分析中，因此，将社会网络映射为一个同异反系统，并采用联系度刻画社会网络研究对象（顶点）间的相似性。该工作重点是如何刻画顶点间的同异反属性。下面通过图 2 所示的社会网络给以说明。

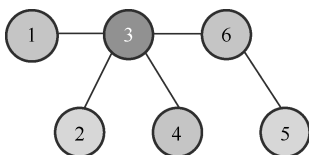


图 2 社会网络

该网络共有 6 个顶点。以顶点对 (1, 4) 为例将其属性记为 $A(1)$ 和 $A(4)$ ，该属性应该包括整个网络研究中的所有顶点，因此，其个数为 6 个。如果将顶点间的公共 1 级邻居作为同属性，则顶点 1 与顶点 4 的公共的 1 级邻居只有顶点 3，其个数等于 1。同理，如果将非共同 1 级邻居作为反属性，则记为 {1, 4, 6}，它包含 3 个元素，数量就是 3。其他顶点为异属性，记为 {2, 5}，一共有 2 个元素。综上所述，顶点 1 与顶点 4 的联系度的定义如下述公式：

$$\rho(1,4) = \frac{1}{6} + \frac{2}{6}i + \frac{3}{6}$$

如果考虑加权网络，定义为顶点间加权联系度，将每一部分都加上权重，如下：

$$\rho(v_k, v_s) = \sum_{l=1}^S w_l + \sum_{l=S+1}^{S+F} w_l + \sum_{l=S+F+1}^N w_l = a_{ks} + b_{ks}i + c_{ks}j$$

由顶点间的联系度可以计算出顶点联系度，即该顶点与所有其他顶点间的联系度的均值，如下：

$$\rho(v_k) = \frac{\sum_{s=1}^n \rho(v_k, v_s)}{n} = a_k + b_k i + c_k j = \frac{\sum_{s=1}^n a(v_k, v_s)}{n} + \frac{\sum_{s=1}^n b(v_k, v_s)}{n} + \frac{\sum_{s=1}^n c(v_k, v_s)}{n}$$

四、 α 关系社区挖掘

基于顶点联系度提出了一种 α 关系社区挖掘算法。其中 α 为给定度量社

区关系的阈值。

1. 最小 α 社区

最小 α 社区的定义如下。

最小 α 社区: $\min SN_{\alpha} = \langle \min relV_{\alpha}, \min relE_{\alpha} \rangle$

最小 α 集对关系结点集: $\min relV_{\alpha} = \{v_k \mid \rho(v_k) = a_k + b_k i + c_k j, a_k \geq \alpha\}$

最小 α 集对关系边集: $\min relE_{\alpha} = \{e_{ks} \mid v_k, v_s \in \min relV, e_{ks} = \langle v_k, v_s \rangle \text{ 存在}\}$

直观来看，在给定的图中，如果其同属性的联系度值大于给定的预值，这些顶点与顶点中连接的边所构成的社区称为最小 α 社区。

通过顶点联系度可以计算出网络中所有顶点的联系度如图 3 所示。

$$\begin{aligned}\rho(1) &= 11/36 + 9/36i + 16/36j = 0.306 + 0.250i + 0.444j \\ \rho(2) &= 11/36 + 9/36i + 16/36j = 0.306 + 0.250i + 0.444j \\ \rho(3) &= 17/36 + 3/36i + 16/36j = 0.472 + 0.083i + 0.445j \\ \rho(4) &= 15/36 + 9/36i + 12/36j = 0.417 + 0.250i + 0.333j \\ \rho(5) &= 10/36 + 8/36i + 18/36j = 0.278 + 0.222i + 0.500j \\ \rho(6) &= 16/36 + 6/36i + 14/36j = 0.444 + 0.167i + 0.389j\end{aligned}$$

图 3 顶点的联系度

图 4 中如果给了 $\alpha=0.4$ ，相同属性中所有满足联系度大于 0.4 的顶点为 3、4、6；这 3 个顶点及其连边一起便是最小 α 社区。

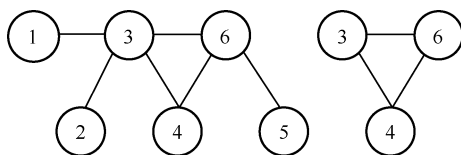


图 4

2. 最大 α 社区

最大 α 社区的定义如下。

最大 α 社区: $\max SN_{\alpha} = \langle \max relV_{\alpha}, \max relE_{\alpha} \rangle$

最大 α 集对关系结点集: $\max relV_{\alpha} = \{v_k \mid \rho(v_k) = a_k + b_k i + c_k j, a_k + b_k i \geq \alpha\}$

最大 α 集对关系边集: $\max relE_{\alpha} = \{e_{ks} \mid v_k, v_s \in \max relV, e_{ks} = \langle v_k, v_s \rangle \text{ 存在}\}$

如果认为所有异属性的顶点转换成同属性，这也是社区中所有可能构成的最大 α 社区，同样仍取 $\alpha=0.4$ ，考虑到同和异两个值加在一起大于 0.4 的，在给定的例子中，6 个顶点都包含在内，所有的边也包含在内。也就是说，对于给定的例子，如果设定 $\alpha=0.4$ ，所有满足最大社区的就是整个给定的图（见图 5）。

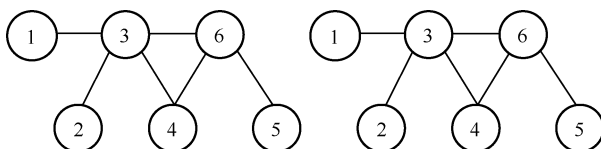


图 5

五、顶点间联系度在链接预测和社区发现中的应用

1. 顶点间联系度模型

目前基于集对理论描述顶点相似性的计算方法如下：

（1）将共同 1 级邻居集定为同，非共同 1 级邻居集定为反，其他的顶点定为异。如图 6 所示，如果考虑顶点 1 和 3，其公共 1 级邻居只有顶点 2 和 4，它们是同。非共同 1 级邻居只有顶点 9 和 10，它们是反。其他的顶点 5、6、7、8 为异。由图 6 可见，反属性顶点 9 和 10 比异属相顶点更容易转换为同，因此该定义不合理。

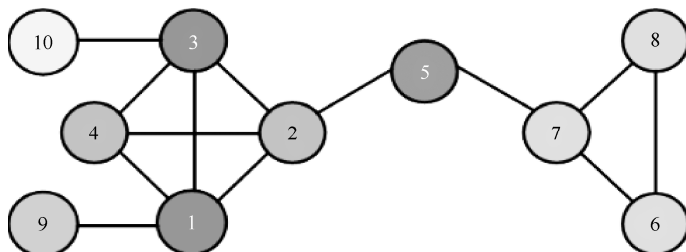


图 6

（2）还有一种定义，将共同的 1 级邻居集表示同，共同的 2 级邻居集表示异，其他的顶点表示反。这里仍以顶点 1 和 3 为例，其共同的 1 级邻居为 2 和 4，共同的 2 级邻居为顶点 5，顶点 5 为异，其他的顶点都是反。上述定义

的问题仍存在，同时该定义还存在一定的问题，如果只观察顶点 1 和顶点 7 之间的联系，它们作为异的顶点，共同的 2 级邻居只有顶点 2，如果观察顶点 1 和顶点 6 只之间的联系，异的顶点只有顶点 5，直观表达顶点 1 和顶点 7 联系的异元素只有一个，顶点 1 和顶点 6 之间的异元素也只有一个，表达的顶点 1、顶点 7 和顶点 6 之间的联系度部分是相同的，但从图中的结构来看，直观感觉顶点 7 应该比顶点 6 与顶点 1 联系更紧密，定义中还未反映出这样的差异。

为此，我们给定了改进的方法，同的定义仍然是共同的 1 级邻居，对于异元素，考虑到共同的 1 级邻居和 2 级邻居之间的关系，图 6 中顶点 1 的 1 级邻居有顶点 2，对顶点 7 而言是其 2 级邻居，顶点 5 是顶点 1 的 2 级邻居、顶点 7 的 1 级邻居，因此，这样将 1 交 2、2 交 1 的元素考虑进来，顶点 1 和顶点 7 之间异的元素就有顶点 2 和顶点 5，顶点 1 和顶点 6 之间只有一个共同的 2 级邻居顶点 5，因此，它们之间异的元素只有一个元素。从这种表达上来看，顶点 1 和顶点 7 之间的联系度比顶点 1 和顶点 6 之间大一些，也能反映出它们在图中的差异性。从而得出新的顶点间联系度的定义如下：

$$u(v_k, v_s) = \frac{(1)_{\text{IS}} \times (w(v_i))_{S \times 1}}{N} + \frac{(w(v_i))_{\text{IF}} \times (i(v_i))_{F \times 1}}{N} + \frac{(1)_{\text{IP}} \times (w(v_i))_{P \times 1}}{N} \times j$$

其中，共同 1 交 2、2 交 1 和 2 级共同邻居作为异属性顶点，分为图 7 所示的几类。

这种定义在链接预测中的应用应考虑如下情况：如图 7 所示，只考虑顶点 k 和顶点 s 之间的联系可能还有不确定（异）的元素都用 i 表示，不确定的元素向确定转换也就考虑这些情况中 i 可能的转换情况。

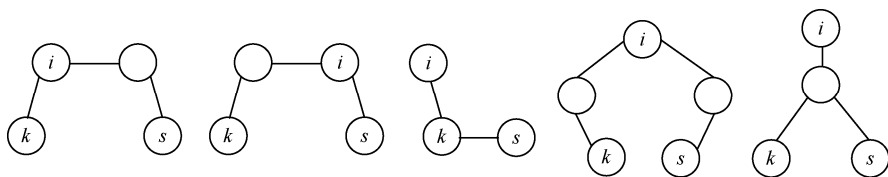


图 7

如图 8 所示，如果只考虑两个红色顶点 k 和 s 之间的关系，它们共同的 1 级邻居只有顶点 1、2、3 为确定属性；其余这种共同的 1 级和 2 级邻居相

交的或 2 级和 1 级邻居相交的顶点为不确定属性。假设顶点 4 与顶点 s 之间在某一时刻建立了联系，顶点 4 就变成了顶点 k 和 s 的公共 1 级邻居，便转换成同了。同理，如果顶点 7 和顶点 k 与 s 分别建立了联系，那么顶点 7 也会转换成共同的 1 级邻居。这就说明顶点的 1 级邻居可以转换成公共的 1 级邻居的条件，与顶点 2 级邻居转换成公共 1 级邻居的条件转换是相同的。如何判断这些顶点之间的转换？为了可以定量描述，定义了顶点的聚集系数，表达式如下：

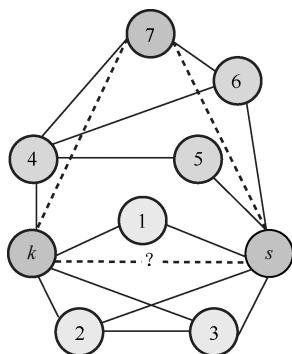


图 8

$$CC(v_k) = \frac{N(v_k)_1 \text{ 中顶点间的实际边数}}{C_{|N(v_k)_1|}^2}$$

为了描述某一个顶点的聚集系数，顶点之间、某一个顶点和其 1 级邻居之间是否会转换成共同的 1 级邻居，所用的方法便是定义顶点的聚集系数，顶点的聚集系数反映与当前顶点相关联的 1 级邻居中，包含的实际边数与所有顶点连接的边数比值，这个比值越大，也就是说 1 级邻居不是顶点的公共 1 级邻居，最有可能转换成它们公共的 1 级邻居。简言之，即异属性最有可能转换成同属性的条件。

如图 9 所示，顶点 2 联系的 1 级邻居有顶点 1、3、4、5，这 4 个顶点之间的连边只有 1、3，1、4 和 3、4 之间右边，其边数一共有 3 个，如果 4 个顶点每两个顶点都有连边，最多有 6 个边，那么顶点 2 的联系度为 $3/6=1/2$ 。

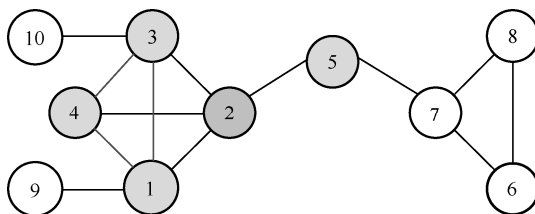


图 9

如果考虑到顶点、边具有权重，如下公式：

$$u(v_k, v_s) = \frac{(1)_{1 \times S} \times (w(v_i))_{S \times 1}}{N} + \frac{(w(v_i))_{1 \times F} \times (i(v_i))_{F \times 1}}{N} + \frac{(1)_{1 \times P} \times (w(v_i))_{P \times 1}}{N} \times j$$

顶点之间的权重量化方法根据传统说法，认为一个顶点对外联系、影响力是确定的，这个顶点的度越多，其对外联系某一个边上的影响力越小。将各个顶点之间的联系度对的影响考虑权重量化，主要借鉴于顶点之间度的关系，基本思想是一个顶点对其他顶点的影响力是固定的，顶点联系度越多，对某一个顶点的影响力越小。换言之，如果一个顶点的影响度是 N ，通过某一个边对其他顶点的影响为 $1/N$ ，这是常规用的办法。借鉴这种办法对顶点之间的联系度和权重进行了量化，根据它们之间联系的路径关系定义了诸多形式。

同理，还有对 i 的量化方法。为简单起见，直接将 i 用顶点聚集系数进行刻画，顶点聚集系数越大，越有可能转换成同。反的元素部分看作 0，不予考虑。

2. 链接预测

图 10 所示为社会网络研究中最典型的例子——空手道俱乐部，这里做链接预测最经典的方法是将其中的 10% 作为测试集，其余的 90% 作为训练集，共有 70 条边，任意选出 7 条边作为测试集，剩下的边构成的图计算所有顶点之间未连边的相似度，计算完成后对其进行评价。主要采用 2 评价指标，分别为 AUC 和 Precision。

(1) 目前最常用的指标之一为 AUC 评价指标，主要衡量整体算法的准确性，定义公式如下：

$$AUC = \frac{n' + 0.5n''}{n}$$

这个衡量指标主要的做法是在测试集中选出一条边，在未知边集中任意选出一条边，如果测试集中的值大于未知边集中的值，那么 $n' + 1$ ，如果两个值是相等的，将 $n'' + 1$ ，这样随机地进行测试实验。

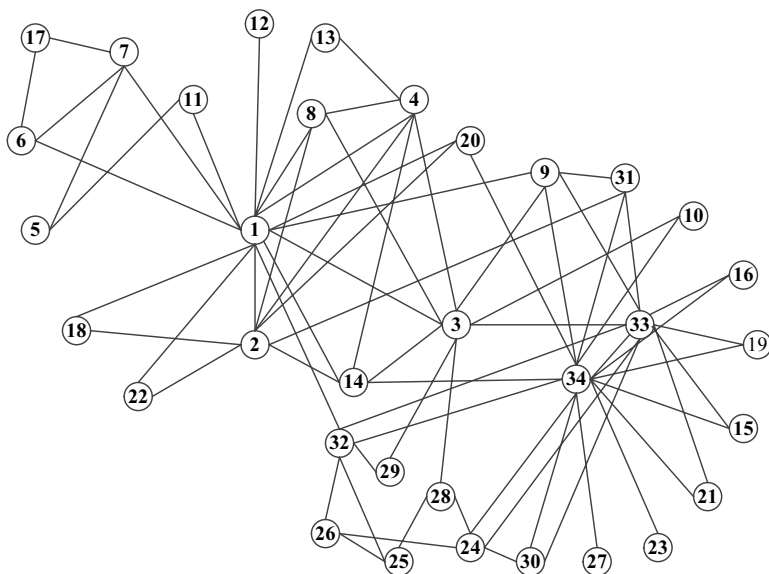


图 10

共同邻居 CN: $S_{ks}^{CN} = |N(v_k) \cap N(v_s)|$

资源配置 RA: $S_{ks}^{RA} = \sum_{v_i \in N(v_k) \cap N(v_s)} \frac{1}{d(v_i)^G}$

Katz 指标: $S_{ks}^{Katz} = \sum_{l=1}^{\infty} \beta^l \times |\text{path}_{ks}^{<l>}| = \beta \mathbf{A}_{ks} + \beta^2 (\mathbf{A}^2)_{ks} + \beta^3 (\mathbf{A}^3)_{ks} + \dots$

基于联系度的顶点间相似性度量指标 SPCD1:

$$S_{ks}^{SPCD1} = \frac{|N(v_k)_1^G \cap N(v_s)_1^G|}{N} + \frac{|V| - |N(v_k)_1^G \cup N(v_s)_1^G|}{N} i + \frac{|N(v_k)_1^G \cup N(v_s)_1^G| - |N(v_k)_1^G \cap N(v_s)_1^G|}{N} j$$

基于联系度的顶点间相似性度量指标 SPCD2:

$$S_{ks}^{SPCD2} = \frac{|N(v_k)_1^G \cap N(v_s)_1^G|}{N} + \frac{|N(v_k)_2^G \cap N(v_s)_2^G|}{N} i + \frac{|V| - |N(v_k)_1^G \cap N(v_s)_1^G| - |N(v_k)_2^G \cap N(v_s)_2^G|}{N} j$$

现有的评价指标有上述 5 种, 做实验主要是与这些指标对比。实验结果如表 1 所示。

表 1 AUC 实验结果 (n=100000)

相似性指标	FB	Jazz	Neural	USAir	Email	NS	PB	PG	Hep
CN	0.8451	0.9385	0.8603	0.9377	0.8525	0.9335	0.9266	0.5923	0.9125
RA	0.8479	0.9520	0.8795	0.9527	0.8542	0.9393	0.9299	0.5919	0.9128
WCCD-S	0.9130	0.9667	0.8606	0.9685	0.8675	0.9438	0.9073	0.5936	0.9128
Katz (4)	0.7460	0.7033	0.7343	0.7787	0.8777	0.9431	0.8348	0.7135	0.9127
Katz (N)	0.8564	0.9560	0.8745	0.9283	0.9245	0.9660	0.9425	0.7256	0.9128
SPCD1	0.8857	0.7877	0.3715	0.4391	0.4851	0.9419	0.5089	0.6879	0.9127
SPCD2	0.7899	0.8314	0.6728	0.8214	0.8601	0.9392	0.8884	0.6517	0.9127
WCCD	0.9250	0.8400	0.9280	0.8710	0.9315	0.9575	0.9315	0.7399	0.9128

表 1 中，只有红框部分是自己的度量方法，数字标红的都是指标较好的。可以看出，有红框的度量方法是自己提出来的，第一行只考虑了同关系，未考虑异关系可能向同关系转换，红字部分表示度量方法结果比较好的。如果只考虑同关系，算法只有在几个网络中没有比现有的度量方法好。第二行算法中，经过实验只考虑了 2 级邻居，即网络中“五度”之内的关系，如果大于“五度”，度量方法的复杂性急剧增加，精度减小。

(2) Precision 评价指标，如下式：

$$\text{Precision} = \frac{m}{L}$$

该指标将所有测试集的边和未知边集的边顶点联系度进行排序，排序后在给定的 L 项中看有多少包含测试集的指标的比例。

实验结果如表 2 和表 3 所示。由此可见，与其他方法相比，我们的方法具有较高的预测准确率。

表 2 Precision 实验结果 (L=100)

相似性指标	FB	Jazz	Neural	USAir	Email	NS	PB	PG	Hep
CN	0.8530	0.8091	0.1947	0.5877	0.3272	0.2871	0.4101	0.1024	0.1025
RA	0.8573	0.8178	0.2921	0.6209	0.2913	0.4307	0.2476	0.0536	0.1538
WCCD-S	0.8566	0.3502	0.1558	0.1652	0.2095	0.5024	0.1762	0.0878	0.1794
Katz (4)	0.7555	0.0221	0.0195	0.0089	0.0269	0.0718	0.0881	0.0488	0.0256
Katz (N)	0.8733	0.7851	0.1947	0.6809	0.3232	0.2871	0.4111	0.0995	0.1025
SPCD1	0.8738	0.5696	0.0292	0.1540	0.0459	0.4307	0.1175	0.0215	0.1538
SPCD2	0.8151	0.3618	0.0292	0.3326	0.0738	0.0000	0.0979	0.0205	0.0769
WCCD	0.9070	0.5773	0.2482	0.1451	0.3991	0.5024	0.4111	0.1170	0.1794

表 3 Precision 实验结果 ($L=1000$)

相似性指标	FB	Jazz	Neural	USAir	Email	NS	PB	PG	Hep
CN	0.0452	0.2126	0.0730	0.1411	0.1144	0.0718	0.2114	0.0334	0.0256
RA	0.0455	0.2379	0.0808	0.1782	0.1139	0.1364	0.1714	0.0183	0.0487
WCCD-S	0.0469	0.2264	0.0584	0.1265	0.0777	0.1364	0.1664	0.0100	0.0487
Katz (4)	0.0247	0.0245	0.0302	0.0077	0.0101	0.0215	0.0666	0.0097	0.0077
Katz (N)	0.0458	0.2078	0.0711	0.1408	0.1125	0.0646	0.2130	0.0360	0.0231
SPCD1	0.0478	0.1163	0.0107	0.0257	0.0078	0.1292	0.0137	0.0077	0.0461
SPCD2	0.0399	0.1232	0.0136	0.0952	0.0384	0.0144	0.0287	0.0125	0.0051
WCCD	0.0510	0.4411	0.0779	0.0324	0.1218	0.1364	0.1781	0.0332	0.0487

3. 社区发现

基于顶点间联系度的定义，提出了社区间联系度的定义，如下：

$$\text{Sim}(C_K, C_S) = \frac{\sum_{s=1}^{|C_S|} \sum_{k=1}^{|C_K|} \mu(v_k, v_s)}{|C_K| \times |C_S|}$$

将顶点间联系度和社区间联系度的度量方法与凝聚型层次聚类算法相结合，提出了一种新的社区发现算法。该算法将联系度高顶点优先聚合形成初始社区，再将具有高联系度的初始社区进行聚合，直到所有顶点聚合为一个社区。然后通过模块度确定社区划分结果，即将最大模块度值所在层作为社区划分的最终结果。

六种度量指标在空手道俱乐部网络中的层次聚类结果如图 11 所示。由此可见，我们提出的方法均取得了最大的模块度值。

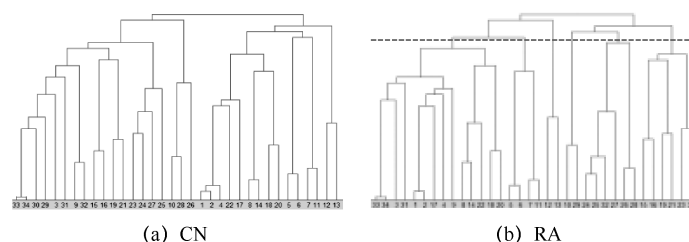


图 11

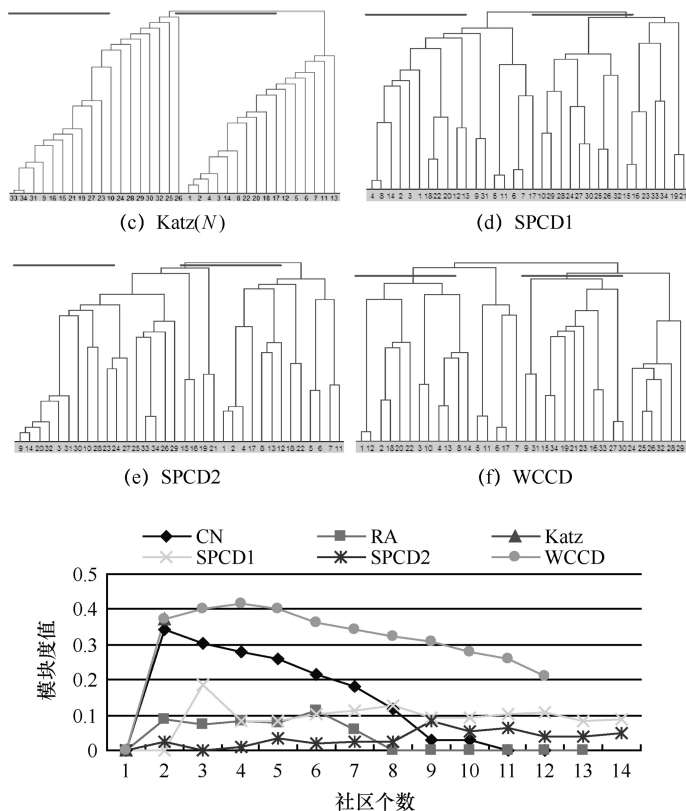


图 11 (续)

同时，又将我们提出的社区发现算法与其他社区发现算法进行了比较，实验结果如表 4 和图 12 所示。可见，我们的算法均具有较好的实验效果。

表 4 五个算法社区划分结果的比较

网络	GN	CNM	LP	SC	VSFCM
Karate	0.401 /5	0.381/3	0.371/3/	0.360/2	0.419 /4
Dolphin	0.519 /5	0.496/4	0.509 /4/	0.394/6	0.519 /4
FB	0.594 /10	0.548/6	0.576/12	0.507/12	0.578 /7
Jazz	0.405 /39	0.439 /4	0.284/2	0.351/8	0.365/4
Neural	0.302/33	0.369 /4	0.322/28	0.103/33	0.348 /3
USAir	0.136/125	0.319 /7	0.001/2	0.267/16	0.328 /13

(续表)

网络	GN	CNM	LP	SC	VSFCM
Email	0.532 /61	0.504 /10	0.014/4	0.412/45	0.474/20
NS	0.958 /91	0.955/276	0.781/38	0.684/33	0.957 /27
PB	0.418/205	0.426/77	0.433/3	0.328/62	0.365/3
PG	0.857/39	0.934 /42	0.871/38	0.830/42	0.931 /38

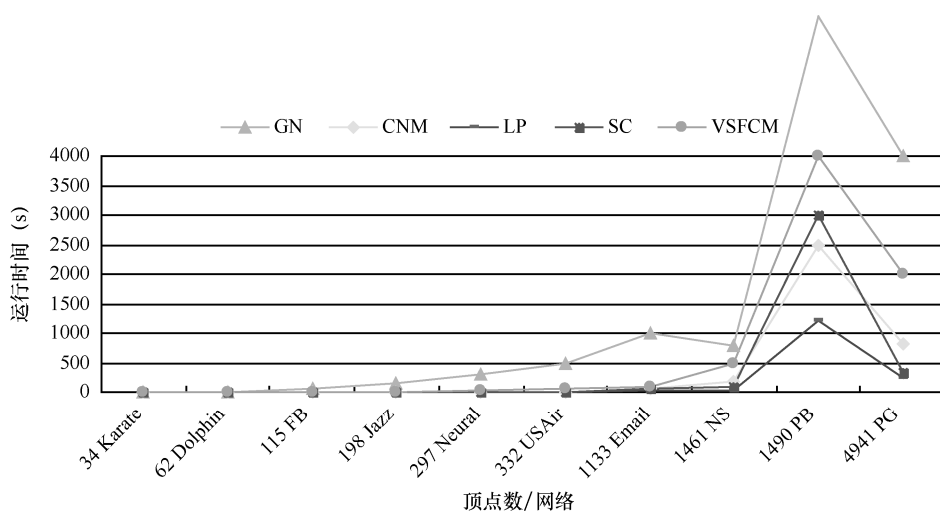


图 12

上述为集对理论中的联系度在传统社会网络中进行链接预测和社区发现的应用。还可以进一步将集对理论应用到其他类型网络中进行相关研究。

作者简介

郭景峰：中国计算机学会数据库专家委员会委员。现任燕山大学信息科学与工程学院计算机科学与技术专业教授，博士生导师。长期从事数据库技术的教学和科研工作，在关系数据挖掘理论、大数据科学、数据密集型计算可视化建模理论等方面有深入的研究和丰硕的成果。

10 大怪象——国外供应链计划和决策类软件在中国的运用

上海大学悉尼工商学院教授 高峻峻

随着“互联网+”的思想和应用在中国快速发展，未来适用于中国企业的供应链计划和决策类软件必然有平台化、大数据化、智能化、自动化、便捷化等特点，并且真正理解中国市场，在中国企业中能够落地实施执行，这样才能够给企业带来价值。

供应链计划和决策类软件，是指帮助企业组织、制定、协同供应链计划，指挥和监督该计划的执行并且衡量企业全面供应链活动的系统。作为一个企业的重要决策系统，供应链计划软件扮演了指挥中枢的角色，并在供应链可视化、仿真化、模型化、数字化过程中发挥了重要作用。

供应链计划和决策类软件主要包括需求计划、库存计划、分销和补货计划、供应计划、物流计划、采购计划、成本效益分析、年度经营计划等基本模块，在零售的业态中还包括商品品类计划、动态定价管理、促销管理等。

目前在中国市场出现的供应链计划软件中，除了少数知名企业如京东、特步等采取了自主开发的模式，所有品牌均来自于欧美。这些国外软件在促进中国供应链管理发展的进程中发挥了重要作用，但也出现了各种水土不服的种种怪象，让很多中国企业对该类软件望而却步。

笔者作为这个领域多年的研究者和实践者，通过走访了大量的企业用户、软件公司、咨询公司等对这些怪象进行了总结和提炼，并分析了问题的成因，目的是让供应链计划软件能够在中国企业中真正生根发芽、枝繁叶茂。

怪象 1：数据清洗——垃圾进就会垃圾出

在使用供应链计划和决策类软件的时候有一句至理名言叫“Garbage in

and garbage out”，原始数据的质量直接关系着这类软件使用的成功与否，因为其核心之一其实就是大数据的分析和挖掘。

然而，绝大部分中国企业的历史数据还存在种种问题，如数据缺失、人为造假、格式不匹配、计算公式不统一、各种活动没有数据记载等。因此，数据的清洗和整理在使用计划软件之前就显得至关重要，然而缺少必要的工具、规则、技术，数据清洗将成为计划人员的一场“噩梦”：1~3 年的历史数据、成百数千的 SKU 数量、纷繁复杂的销售渠道、眼花缭乱的市场活动等无不给计划人员带来巨大的挑战。

现有软件的数据清洗功能还远远不能给计划人员提供支持甚至是处在缺失状态，往往需要计划人员在线下对数据进行人工手动清洗，然后批量导入到软件中去，这就给数据清洗工作带来了很大的难度，很大程度上打消了计划人员对数据清洗的工作热情，不得不少清洗、部分清洗甚至是不清洗，这些都给该类软件的日后使用带来了巨大的隐患。

怪象 2：黑箱操作——神奇的水晶球

曾经有一个中国企业使用了一款国际非常知名的此类软件，该软件号称能提供供应链计划领域的全球“最佳实践”，其运算公式、算法、逻辑等都实施了黑箱操作，即该软件会自动运算，然后告诉计划人员运算的结果，而不会解释任何原因。这就导致了计划人员完全不知道这些结果是怎么计算出来的，即使自己非常清楚关于市场的最新动态和情报，也无法对这些结果在系统中进行任何调整和修正。

同时计划人员也无法向其他职能的同事和领导层去解释软件运算出来的结果，因为根本就不知道逻辑和算法，这就把计划人员放在了一个很尴尬的位置，完全体现不出自己价值和意义。

怪象 3：汇报和展示——计划的价值在哪里？

有一位非常资深的供应链计划的职业经理人曾经说过，供应链计划软件的汇报和展示功能其实不亚于其数据分析功能。

计划工作特别是供应链计划是整个企业端到端供应链驱动的核心。从客

户的需求计划做起，到整个销售渠道中的补货和调拨，再到后端供应商的物料供应计划，以及中间的生产计划和排程，这些无不需要计划职能的指挥、协调、监控等。然而，计划职能的工作往往是无形的和幕后的，并且需要和其他各个职能如销售、市场、财务、研发、生产等进行深入而广泛的协同和沟通。

因此，计划软件的汇报和展示功能对于计划人员就显得至关重要，计划软件的该类功能要让其他职能的同事特别是领导层迅速、清楚而又深入地理解计划工作的重要性，以及日常计划工作的决策及背后的逻辑和原因。

计划工作的价值和意义假如不能被理解的话，被裁撤的风险就大了。然而，现有不少软件的该类功能往往不是其核心，还局限在只是能够生成简单报表的层次上，并且界面复杂，计划决策的逻辑和过程不能很好地展示出来，导致计划人员被其他职能的人认为只是“搜集数据的”或者“根本不知道他们在说什么”的尴尬局面出现。

怪象 4：缺少针对性——外国的月亮一定比中国的月亮圆吗？

目前市场上的该类软件都是来自欧美的舶来品，均是针对欧美成熟市场的特点、发达的供应链管理体系、较高的操作人员的素养，并基于经典供应链管理理论于 20 世纪 90 年代所开发的产品，对于理性消费、数据质量高、产品生命周期长的成熟市场发挥了很强的效用。

然而，这类软件对于中国的市场和企业的特点普遍缺少针对性研究，例如销售渠道漫长复杂且不透明、市场竞争异常激烈、产品生命周期短、基础数据质量很差、使用人员经验欠缺、没有完善的供应链管理体系等，特别是目前在中国蓬勃发展的新零售、电商、时尚类企业，其独特的销售和供应链管理如商品管理、协同定价、快速反应等特点更是没有被该类软件所关注。

生搬硬套此类软件的中国企业在付出惨重代价后，最终发现外国的月亮未必比中国的月亮圆。

怪象 5：客户体验——体会到我的情绪了吗？

供应链计划和决策类软件，除了稳定和功能强大，用户体验也很重要，因为该类软件的逻辑普遍较为复杂，如果用户体验差了，会严重影响这类软件的使用效果。当前该类软件的开发人员和测试人员往往都具有 IT 背景，非常强调其功能和性能，特别是大量按键和数据信息在同一界面上的展示，忽视了软件和使用人交互的重要性，让使用者感到功能不容易理解、信息检索普遍困难、眼睛容易疲劳、无法向其他人展示等。

特别是随着互联网和手机应用在中国的迅速发展，中国使用者对于软件的要求也越来越高：极简、便捷、人性化、尽量不用说明书、能点一键就不点两键。所以，只有撇去开发者和测试者的身份，以用户的角度去审视软件，不断改良，才能让软件具有更好的用户体验。

怪象 6：数学模型——我们不是陈景润

此类软件的一个重要功能是数据分析和挖掘，必然会涉及复杂的数学模型的选择和建立，以及大量的计算，这些都是该类软件成功运行的基础之一。

然而，不少现有软件太过于偏重数学和统计学模型，把该软件完全视为一个专业的数据分析型工具，没有把这些模型和中国企业现有业务模式和流程进行对接和融合，忽视了该软件在企业中操作人员对于数学和统计学的理解和应用水平，没有把这些数学模型和业务逻辑结合在一起，导致操作人员畏惧使用该类软件，不能对该软件的功能进一步深挖。

曾经有一名供应链计划经理说道：对于计划软件，我就如同一名司机，要熟练掌握操作技能，并能够给同事和领导汇报清楚，但对于背后的运算逻辑、数学模型参数的配置等细节问题我只需要了解就行，因为其他人不理解，也不关心。

怪象 7：重软件轻管理——手术很成功，但病人却死了

供应链计划和决策类软件公司往往过分宣传自己产品的功能，以及所能

给企业带来的价值，把自己的工具打造成了一个万能药，而忽略了供应链管理“内功”的提升。

中国企业供应链水平的提升是“人、流程、工具”共同发展的一个综合体，特别是使用计划类软件人员的职位，一定要事先确立，并且使用者的素养和能力要达到一定的水平，而支撑这类软件顺利使用的供应链计划流程、职责、汇报关系、组织架构、绩效考核等也要能够同步发展。

但是不少软件公司由于完全的销售目标导向，往往有意或者无意忽略以上因素，对于供应链管理基础的改善往往少提或只字不提，甚至有时会堂而皇之地告诉企业要依靠软件来倒逼供应链流程的改善。

软件公司所追求的就是软件上线，至于上线以后谁来用、用的怎么样、是否和流程匹配则不愿提及。因此，“手术很成功，但病人却死了”，系统上线成功，但没用起来的怪象比比皆是。

怪象 8：过分宣传——下一个世界 500 强是你吗？

大部分该类软件都来自国际知名软件公司，不可否认的是不少国际知名品牌公司都使用了这些软件公司的产品。因此，这些软件公司在中国的宣传未免有过分、拔高、偷梁换柱之嫌。

例如，软件公司会把该类软件打造成一个“万能药”，目前中国企业所遇到的预测不准、交付率低、库存高企、效率低下、浪费严重等问题，似乎上了这一个计划软件就能包治百病，似乎造成这些问题的主要原因就是缺少这样一个软件。

还有软件公司宣传世界 500 强中有不少公司都在用他们的产品，给中国企业的感觉就是如果上了类似的软件就成了世界 500 强一样。在这种情况下，所谓的客户，就是该类软件的买单者往往会成为他们的主要进攻目标，前景渲染、成功案例、战略发展等都在笼罩在这类软件的价值之中，而忽略了用户的体验和效果，以及这类软件所能够带来的真正价值。

怪象 9：售后服务——一个无奈的差评

该类软件由于其特有的逻辑复杂性和高度的业务匹配度，在软件上市后需

要持续和专业的售后服务，特别是有经验的咨询和实施顾问进行深入的辅导来解决用户的问题，例如，清除 Bug、挖掘新的功能、更加贴近业务，然而现有的软件供应商受制于顾问资源、行业经验、咨询成本等因素，往往对客户的要求不能很及时地进行反馈，导致简单的问题解决周期较长，且功能挖掘不完善。

例如，某用户受制于对参数配置和理解不透彻，导致某一功能一直没有启用，而该软件的供应商对此问题带搭不理，最后不得不动用公司高层关系从国外请来一名咨询顾问来解决，耗时耗财不说，该外籍顾问说这一类简单问题其实根本不需要他来。

怪象 10：价格和周期——有钱就任性吗？

这个话题其实也是老生常谈了，这类软件的价格往往不菲，动辄几百万元甚至数千万元，后续还有大量的维护和升级费用，并且实施周期取决于使用的模块的多少，一般都在几个月甚至几年时间。在实施的时候需要大量的顾问团队、实施团队、客户使用团队人员的介入，并且涉及企业各方面的职能，这些都给很多中国企业带来了很大的财务和心理负担，也给上线不成功带来了很大的风险，这些因素交织在一起，往往让中国企业很难下定决心去使用这类软件，难以量化的收益更是让客户举棋不定。

当然，造成以上怪象的责任也不全在软件公司，中国企业也要认真思考如何去从“人、流程、工具”等角度全面进行供应链改善和提升。随着“互联网+”的思想和应用在中国快速发展，未来适用于中国企业的供应链计划和决策类软件必然有平台化、大数据化、智能化、自动化、便捷化等特点，并且真正理解中国市场、企业和使用者，真正在中国企业中能够落地实施执行，这样才能够给企业带来价值。

作者简介

高峻峻：工商管理系主任、上海大学需求链研究院院长、博士生导师、教授。研究方向：大数据需求预测、电商供应链、时尚品供应链。主讲课程：电商供应链需求管理——预测和大数据应用，动态库存——基于 IBP 模拟的优化分析。

大数据和人工智能在高校舆情处理中的应用

西南科技大学教授 张 晖

目前，计算机已非常普及。据统计，西南科技大学 70% 的学生每天上网的时间达到 2~5 小时，30% 的学生的上网时间达到 5 小时以上。在这种情况下舆情一旦发生并快速传播，将会对学校造成无法挽回的损失。

西南科技大学在舆情管理方面有 3 支队伍：第一，党委宣传部中有一个网络管理科，负责网络舆情的内容审查，以及学校官方微博、微信的维护；第二，团委中有一个队伍，发现舆情后，进行正面宣传；第三，就是自行开发的信息化系统，负责网络舆情的自动发现及学校舆情的自动监控。

一、舆情系统架构

系统的架构大概分为三个部分。

（一）系统逻辑结构

图 1 所示为系统逻辑结构。

数据采集层采集的数据对象有新闻、网页信息、BBS 信息及微博信息等，采集信息使用了网络爬虫技术，将信息采集后放在服务器上。由于所采集的网站多、信息量大，我们使用了分布式爬虫技术与大数据技术，通过一个刀片服务器中的十个服务器同时进行采集，采集后放到学校的大型存储器上。对存储的数据进行数据清洗和抽取后，存入舆情数据库，再进行多维数据分析、情感分析、热点分析等分析工作，分析后的数据将形成舆情报告，并发送到微信上，为宣传部的工作提供指导。

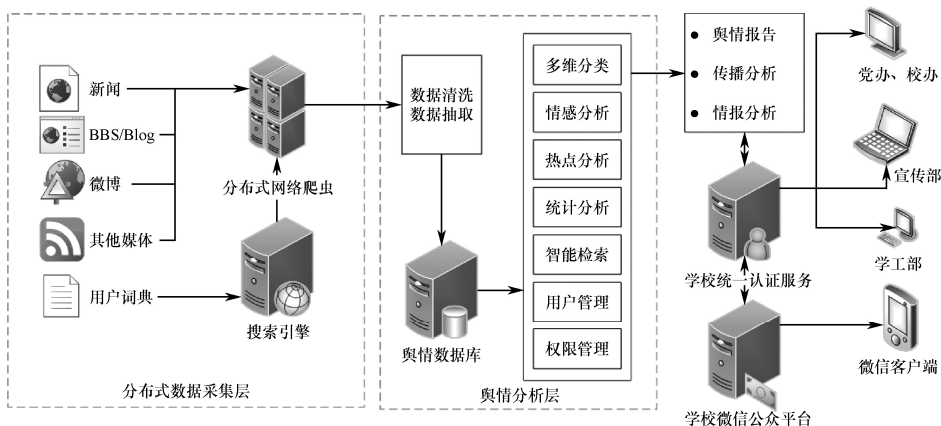


图1 系统逻辑结构

(二) 系统业务结构

图2所示为系统业务结构。

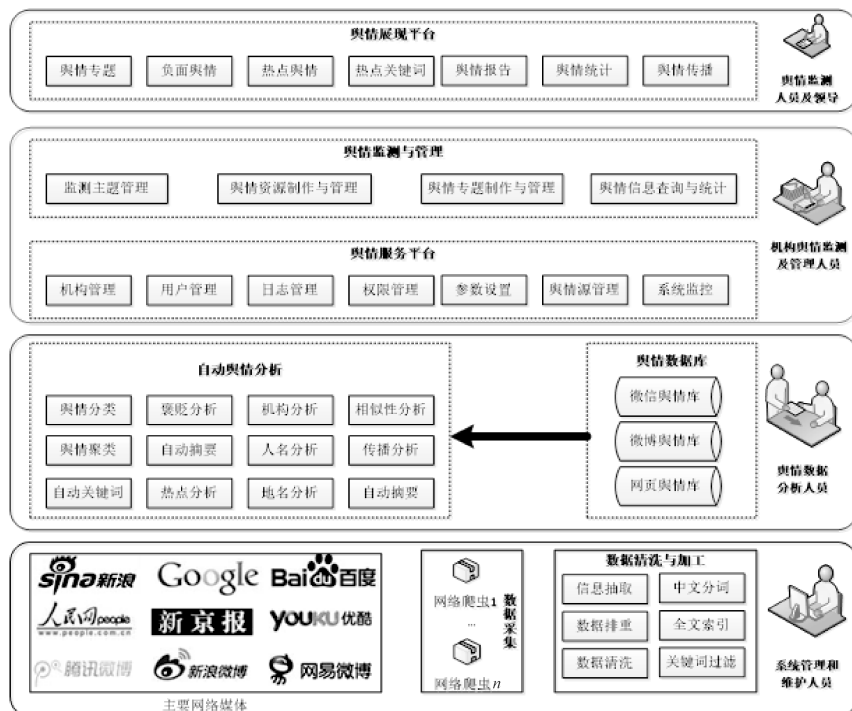


图2 系统业务结构

在基础架构上，主要使用了以下几种平台：第一，大数据技术平台。上述提到的分布式爬虫技术综合应用了 Redis 和 MongoDB 进行数据存储，Redis 快速将爬虫爬到的数据进行预处理，处理好后的数据存入到 MongoDB 中。中间的一层是基于统计机器学习和复杂网络理论的舆情处理层，主要技术是话题的演化（对已发生话题、如何发展、正面的话题、负面的话题信息进行分析），分析后的数据在一段时间后需形成摘要，如一周的舆情摘报，传给宣传部整理后发送给相关部门，在此过程中也需要个性化推荐技术。对外发布基于 SaaS 的云服务，在服务器上发布后，其他高校的宣传部无须安装服务器，只需一个账号即可直接看到其舆情信息。

（三）分布式网络数据采集

目前的分布式爬虫技术已做到近 200 个网站的实时监测，通过基于大数据的分布式采集，其响应时间为 3~5 分钟，如果网页上有变化，3~5 分钟便可察觉，每天的信息增量为 8000~10000 条。除爬虫之外，为方便了解其他搜索引擎所用到的一些搜索结果，也使用了元素引擎，将学校的主题发送到百度等搜索引擎中，将搜索出的结果与爬虫搜索的结果进行组合使用。

舆情系统的功能及核心技术便是分层的处理结构。在信息获取到后，关键在于舆情分析，具体分为三层：最简单的便是用户所设定的关键词，如学校、学院、校领导的名字等浅层信息。接着在关键词和本体上计算与主题的相关性，具体是以主题模型来实现，将一周的舆情做一个文本摘要提供给宣传部使用。通过这些摘要可以发现学校的活跃话题，以及发现关于舆情的发展是否会形成某一个话题，是否最终会变为一个舆情。最后，需要对舆情的正负面进行分析。

图 3 所示为早期实现的系统 4.0，其中虽未应用到大数据处理技术，但用了人工智能技术，当时的爬虫是基于单机做的一个效果。

图 4 所示为目前做的系统 5.0，其中应用了大数据处理技术，已达到比较好的性能，满足了学校对于舆情处理的需要。

赋能大数据教育

全国高校大数据教育教学经验谈



图3 系统4.0

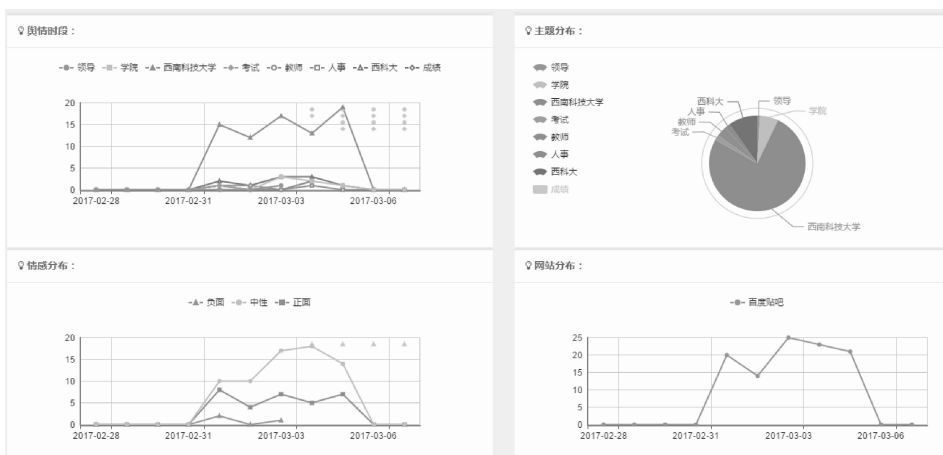


图4 系统5.0

二、舆情系统的特点

- (1) 爬虫 7×24 小时不停监测与学校相关的信息，响应时间能达到 3~5 分钟。
- (2) 97%以上的信息抽取准确率，提高了舆情发现的精度。
- (3) 基于机器学习的自然语言处理，深入分析舆情倾向、热点、趋势等信息。
- (4) 多维度统计分析，深入理解舆情的传播过程。
- (5) 整合高校微信公众平台，让相关工作人员及时掌握舆情动态。
- (6) 采用云计算模式，监测信息及时，节约用户开支。

作者简介

张晖：中共党员，博士，教授，硕士生导师，现任网络信息中心主任。1993 年本科毕业于四川建材学院工业电气自动化专业，2000 年硕士毕业于武汉理工大学机械电子工程专业，2006 年博士毕业于日本北陆先端科技大学院大学（Japan Advanced Institute of Science and Technology）知识科学专业。历任西南科技大学计算机科学与技术学院软件工程教研室主任，计算机科学与技术学院副院长，网络信息中心副主任，网络信息中心主任，现任理学院党委书记。

物联网与农业大数据

华北理工大学数据科学实验中心主任 陈学斌

物联网的理论模型主要分为 3 个层次：最底层是泛在化的传感器网络，中间层是异构性的互联网基础设施，最上层是普适性的数据分析服务。对于物联网的研究，一般也是从这 3 个层次开始的，最底层的感知、中间层的互联互通及最上层的智能化处理。从这个角度来看，物联网的关键技术主要在于 3 个方面：传感器、网络互联和智能信息的处理。

一、物联网应用存在的问题

近年来，物联网在我国应用的领域越来越广泛，也暴露了一些问题。

（一）技术标准问题

世界各国在物联网的研究方面都出台了不同的标准，而这些不同的标准必然会导致传感器信息在互联互通方面存在问题。

（二）安全问题

由于传感器的大量使用，使得信息的采集频率越来越频繁，数据采集的内容越来越广泛，数据的安全也是物联网中必须考虑的一个重要方面。

（三）协议问题

物联网是互联网的延伸，在物联网的核心层面是基于 TCP/IP 的，但在接入层面协议类别就五花八门了，因此，物联网需要一个统一的协议。

（四）IP 地址问题

每个物品在物联网中都需要被寻址，便需要一个地址，物联网中需要更

多的 IP 地址，IPv4 的资源已经耗尽，就需要 IPv6 来支撑，对现有设备来说，如何从 IPv4 向 IPv6 过渡将是一个艰巨的过程。

（五）终端问题

物联网终端除具有本身功能外，还拥有传感器和网络接入等功能，不同行业的需求千差万别，如何满足终端产品的多样化需求，对运营商而言是一大挑战。

随着物联网的应用越来越广泛，物联网采集的数据量不断增加，对于物联网采集数据的如何处理将是未来重点研究的内容。

二、农业大数据的应用

大数据理论最早启蒙于信息资源管理系统，信息资源管理是 20 世纪 70 年代末在美国出现的一个新的概念。大数据的意义不仅在于容量之大，更多的寓意在于人类用于分析和使用的数据在大量增加，通过数据的交换、整合、分析，可以发现新的知识，创造新的价值。传统的数据主要来自业务运营支撑系统、企业信息系统，而当前爆炸式增长的新数据更多来源于物联网、互联网，而物联网所采集到的数据是大数据处理中一个重要的组成部分。随着数据中非结构化数据的占有量不断增加，对于数据处理的方法也提出了更高的要求。目前，整个社会对于数据的认知都发生了改变，大数据让社会各个层面都开始认识到其重要性。数据将成为比黄金、石油更有价值的战略资源。但数据要素和传统的劳动力资本的物质要素在技术特征和经济特征方面还是有一些区别的。

在现代农业中，随着传感器、智能移动设备及互联网的发展，数据也呈现了爆炸式的增长，经济的数据也日益重要。在农业领域，由于地域性、季节性、多样性和周期性的差异，导致农业领域中的数据和互联网中的数据还是有着本质的区别的。在农业中，每年产生的数据量约为 8000PB，其中包括农业自然资源的数据、农业生产的数据、农业市场的数据及农业管理的数据，而未来农业数据还要以每年 50%~80% 的速度增长。到 2020 年，农业大数据应用的市场规模将达到 250 亿~300 亿美元。

三、农业大数据应用带来的挑战

农业大数据的发展也对传统的数据处理技术、体系提出了一个巨大的挑战，需要在数据的采集、标准、处理、分析、展现等方面做一个全新的数据升级。

数据采集。数据采集过程中，由于农业的特点，在采集方面更多地依赖于传感器，包括可植入、可嵌入式的数据获取技术，微型移动信息获取技术，以及生物传感、微纳米传感器、便携式传感器等新型设备，将在农业物联网中得到更广泛的应用，这些设备如何更好地在农作物生长环境中长期有效地采集数据将是一个重点研究领域。

数据标准的统一。数据的增值关键在于整合，而整合的前提就是数据标准的统一。因此，在采集、传输、存储、整合四个方面都需要制定相应的标准。

数据处理技术。这方面需要更多地建立模型，如农作物生长与产量形成积累的模型、农产品消费行为与消费量变化动态的模型、基于多代理系统进行农业智能仿真的模型，以及农业生长环境要素与农业灾害预防之间的模型。

数据展现技术。在大数据背景下，在交互式数据可视化技术的支撑下，我们需要对农业大数据针对不同的阶层、不同的用户，采用不同的展示方式，以方便用户更好、更快捷地使用农业大数据分析得到的成果。

可以说，大数据技术与农业领域的深度耦合，将对我国的农业市场、监测预警、智慧农业生产管理、农业国家宏观管理决策及农业灾害预防等方面都带来前所未有的变化。利用物联网进行农业生产过程监控是现代农业稳定发展的一个重要基础之一，尤其是对大型生产的农作物长势长相监测、灾害预测等。我们的团队在农业灾害预警领域做了一些工作，近些年我们在这一领域申请了两个科技公关项目，主要针对小麦种植过程中的灾害发生进行预测，在前端主要是采集小麦农田环境的一些要素，包括土壤、空气、温湿度及施肥量。还通过视频对小麦的长势进行采集，最终将这些数据汇总到数据中心。在数据中心设计了算法、搭建了一个数据分析平台，对采集的数据进行综合分析。最后结合专家系统对灾害的发生进行预测。目前，这一成果拿

了两项科技进步奖。

对于农业大数据的应用，不仅局限于灾害预测这一领域，未来针对农产品市场出现的区域性、季节性、结构性的卖难买贵问题，利用农业大数据进行优化，实现产销精准匹配，也将是一个重要的研究领域。为了使农业大数据得到更广泛的应用，未来农业大数据应该在对内共享和对外开放两个方面做更多的工作。进一步完善农业大数据生态体系结构模型，包括建立农业大数据资源库、农业大数据云中心、农业大数据采集网，在此基础上建设数据交易平台、数据挖掘分析平台，从而使农业大数据得到更广泛的应用。未来农业大数据的典型应用场景将会越来越多。

在大数据时代，有这么一句话：万物皆比特，一切皆数据，数据的共享、开发、融合将会产生核聚变，大数据的汇聚、治理、应用能力也将会成为现代农业发展的核心力量。

作者简介

陈学斌：工学博士，教授，硕士生导师。现任华北理工大学数据科学实验中心主任，唐山市数据科学重点实验室负责人，河北省数据科学与应用重点实验室负责人，中国计算机学会理事、高级会员，中国计算机学会计算机应用专业委员会秘书长，中国计算机学会高性能计算专委会委员，大数据专家委员会委员。主要研究方向为网格计算、云计算、大数据、网络安全。



未 来 篇

大数据带来的机遇与挑战

大数据时代：机遇、挑战与思考

东北大学软件学院院长 王兴伟

“大数据”一词最早出现于 1980 年，它是由著名的未来学家阿尔文·托夫勒在自己的《第三次浪潮》一书中提出来的。当时的阿尔文·托夫勒将大数据热情地赞颂为“第三次浪潮的华彩乐章”。大约到了 2009 年，“大数据”这一词才成为信息领域的流行词汇。据悉，互联网上的数据每年大约会增长 40%，每两年便会翻一番，目前世界上 90% 的数据是在近几年才产生的。2010 年以互联网为基础所产生的数据比之前所有年份的数据总和还多，不仅数据量在激增，而且数据结构也在发生演变。

一、大数据的特点

除数据激增、数据结构演变等最基本的特点外，大数据还具有如下特点：变化性，即数据随着事物的变化而变化；真伪性，我们获得的数据不完全是可信的，有可能是有问题的，收集到的数据并不是我们所预期的，也不是我们所需要的；价值易逝性，数据的价值有一定的时效，如股市数据；还存在数据对事故的过量描述，或对同一事故不同的表达方式；易损性，数据有可能丢失，也可能被篡改；可视化，数据需要以合理的形式呈现给用户，加以适当表达，否则，绝大多数用户对数据是无法理解的；可验证性，含有数据的分析结果必须是可以验证的，得以验证其正确与否。

二、大数据提供怎样的机遇与挑战

大数据是继云计算、物联网后 IT 领域的又一次重大变革，它提供了难得的洗牌机会，数据已成为国家的宝贵财富、单位的核心资产，影响极其深远，并不局限于 IT 领域，实际上它对国家的治理模式、企事业单位的决策组织业

务流程、个人的生活方式都可能会产生很大的影响。如致力于管理的层级更加扁平化，网民和消费者的界限正在逐渐消失，过去网民是网民，消费者是消费者，而如今网民在浏览的同时便可以完成消费行为。另外，充分利用大数据高效地分析信息，便可以准确地捕捉需求，满足受众的需要。例如，每天收到的垃圾广告，如果通过数据对网民上网习惯进行分析，便可以推送相应的广告。单位的界限也变得模糊，业务模式、文化和组织都需要进行重构。

关于大数据技术，需要关注的是如何解决在大数据收集、处理、存储、传送、管理和维护过程中需要突破和发现的关键技术。再次，是大数据工程，我们需要构建大数据的规划、运营和管理的系统工程。最后，其实也是最重要的是开发大数据应用，基于大数据解决现实问题才是根本所在。发展大数据产业的基础已经具备，即已经有了无所不在的信息聚集与积累，以及信息采集、存储、分析和挖掘技术。发展大数据产业的动力也已经形成，遍布世界采集不尽的信息资源为发展大数据产业提供了可能。看似毫无意义的信息，经过技术处理可以转换成有交换价值和使用价值的信息，这为企业创造财富提供了可能。发展大数据产业迫在眉睫。从狭义的角度来看，大数据产业就是发展与大数据的采集、处理、存储、管理、运行、维护等相关的 IT 产业，要开发相应的软硬件设备，建设大数据中心等。从广义的角度来看，发展大数据产业，就是要与消费、生产领域相融合，发展大数据应用产业，如医疗健康大数据、交通数据、运输大数据等。与此同时，培育大数据相关人才也刻不容缓。麦肯锡预测，到 2018 年，仅美国就需要深度数据分析人才 44 万到 49 万人，缺口达 14 万到 19 万人，需要既熟悉本单位的业务需求又了解大数据技术与应用的人才 150 万人，我国的需求一定会更大。

为此，中国已将大数据技术与产业优先发展列为战略新兴产业。国务院于 2015 年 5 月 8 日发布了《中国制造 2025》，明确提出要加强高端服务器大容量存储设施的建设，促进云计算和大数据应用。2015 年 7 月 1 日发布的《推进“互联网+”行动指导意见》也明确了大数据设施的建设目标，提出了鼓励大数据应用、大数据技术开发的政策，承诺推动公共数据资源开放。2015 年 8 月 9 日，国务院又通过了《关于促进大数据发展的行动纲要》（以下简称《纲要》），在《纲要》中明确指出要推动政府信息系统和公共数据的互联互通，消除信息孤岛，整合各类政府信息平台，避免重复建设，增强政府的公信力，

促进社会信用体系建设。国家要推动交通、医疗、就业、社保等民生领域，政府数据向社会开放，在城市建设、社会救助、质量安全、社区服务等方面开展大数据应用与示范，提高社区的治理水平。

国家要顺应潮流，引导支持大数据产业发展，以企业为主体、以市场为导向，加大政策支持，着力营造宽松、公平的竞争环境，建立市场化应用机制，深化大数据在各个行业的创新应用，催生新业态、新模式，形成与需求紧密结合的大数据产品体系，使开放的大数据成为促进创新创业的新动力，强化信息安全保障，完善产业标准体系，依法依规打击数据滥用、侵犯隐私等行为，让各类主体公平分享大数据带来的技术、制度和创新的红利。

三、大数据带来的启示与思考

第一是研究的超前性。其实早在 1980 年托夫勒便指出大数据的极端重要性和美好未来。假设我们早在 1980 年就意识到大数据的极端重要性，超前部署相关研发工作，那么目前的局面又会如何？因此，我们在寻求引领发展和先发优势时，它呼唤着极具前瞻眼光的战略思维和视野，这也告诉高等学校必须做超前的人才培养，必须做明天的科学研究，企业必须有超前的技术储备，为明天产品的开发做积累，如华为等智能手机的成功都不是偶然的，就在于其前瞻性的布局。

第二是人才培养与知识体系的快速适应性。面向未来的人力资源需求，如今的 IT 类专业人才的培养模式、知识体系、实验课程设置应如何应对？如何培养复合型的人才等都是高等教育领域需要考虑的问题。个人认为，不宜过多设立过度的大数据专业与学院。从狭义的大数据角度来看，更多的是要将大数据作为计算机和软件工程等专业方向，面向大数据对 IT 提出的新要求，如计算机系统结构、硬件结构设计、网络体系结构、软件体系结构、编程模式等，以及数据采集、存储、组织、访问管理等。从广义大数据的角度来看，更多的应与其他专业相交叉，培育大数据方向的复合人才，如金融大数据、工业大数据、金融大数据、交通大数据等，这样才能更好地适应大数据在行业发展的需要。因此，高等教育必须迅速适应大数据要求，适时进行专业培养方向转型，这样便不会出现几年后面对巨大的人才需求所谓的人才过剩及

人才匮乏等现象。在人才培养过程中切忌跟风，不要盲目蜂拥而上，必须看自己是否有真正意义上的教师资源，否则也不会达到很好的效果。

第三是加快发展大数据产业。一个产业的兴旺发达必须具备以下两点：一是市场需求，没有市场需求的产业是不可能长久的；二是与产业相关的各个主体的利益保障，主体权益得不到保障，便没有发展产业的动力。在完善大数据产业政策的过程中必须考虑解决如下问题：社会成员隐私权的保护、社会各类主体信息所有权的保护、公共数据资源开放与国家安全的甄别。思考这些问题不仅是政策制定者的责任，也是大数据技术开发工作者的责任，在数据进行专业化处理的过程中必须兼顾隐私的保护和数据资源公共安全的保护。在大数据产业发展的过程中必须科学布局、理性发展，切忌低水平地重复建设。如今到处在建大数据中心、智慧城市，但这些并不等于大数据产业中的全部。我们需要适当聚焦基础理论、关键技术及核心软硬件产品的突破，需要进一步梳理需求、开发大数据应用，这样才能使大数据落地生根，需要选择应用发展，这样才能使得大数据在中国更好地普及。

在此过程中一定要注重与其他关键技术领域的协同发展。如与云计算、物联网等的协同发展，云计算本身可以为大数据提供解决方案，物联网是数据采集和获取的主要手段。在发展大数据时要注重与国家战略的联动统筹，如与互联网+、中国制造 2025 等的联动统筹，这样才能更好地推动中国社会经济发展的提效。另外，随着数据处理技术的发展，特别是图像语音等分析处理技术，大数据已不像如今理解的那么大。大数据量很大，但价值不一。真正有用的数据是呈现某种状况，如安全事件在几秒时间的影响。如今无法实时理解和分析发生事件的信息。因此，随着数据处理技术的发展，大数据可能没有今天想象的那么大了。

作者简介

王兴伟：博士，教授，博士生导师，教育部新世纪优秀人才，辽宁省优秀教师，教育部创新团队成员；东北大学学科建设与发展处处长，高等教育研究所所长，211 工程办公室主任，985 工程办公室主任；教育部高校本科教学工作水平评估专家；中国教育和科研计算机网 CERNET 专家委员会委员，

CERNET 东北地区网络中心主任；中国教育科研网格 ChinaGrid 专家组成员；中国计算机学会体系结构专委会委员；中国计算机学会高性能计算专委会委员；中国计算机学会高级会员；辽宁省互联网协会常务理事；《东北大学学报》编委；《中国教育网络》编委；曾参加日本名古屋大学校际交流，中国香港理工大学访问教授。一直从事下一代互联网、自组织网络、IP/DWDM 光 Internet、移动无线 Internet、网络信息安全和网格计算等方面的研究工作。

互联网+大数据=大智慧

厦门大学信息科学与技术学院计算机系副主任 张德富

一、大数据时代来临

随着移动互联网和物联网的发展，人与设备均有机连接在一起，通过移动互联网和物联网，人们可以采集到很多数据，其中包括结构化数据、半结构化数据及非结构化数据。大数据主要由非结构化数据构成。

（一）大数据之“大”

众所周知，我们每天都面临着大量的数据，无论是 Internet 网络传输的电子邮件，还是 Facebook、微信、QQ 等传播的信息。2020 年，全球数据总量将达到 40ZB，将 1ZB 数据用 A4 纸打印出来，堆叠起来相当于 585 栋帝国大厦。通常所说的海量数据的规模，指的是 1TB 以上的数据。

（二）大数据的概念

大数据是指数量大、变化快和多样化的信息资产，需要快速地处理，从而给出决策，促进洞察力及优化流程。大数据有不同的定义，但其意义在于通过对海量数据的获取、整合、分析发现新的知识，创造新的价值，带来知识和大智慧。

（三）大数据的“4V”特性

Volume（体量）——通常指数据量的大小，随着计算机技术的发展，存储非结构化数据（如视频、图像数据）变得很容易。一般来说，非结构化数据占数据总量的 80%以上。

Variety（多样性）——由于数据采集的方式多种多样，特别是数据的格式

也各不相同，如图像、视频、语音等数据都具有异构和多样性。

Value（价值密度）——挖掘大数据的价值类似于沙里淘金，从海量数据中挖掘稀疏但珍贵的信息，是一种挑战。

Velocity（速度）——要实时处理，如 1 秒内给出答案。

（四）大数据带来的挑战

1. 从数据的不同维度看待数据会得到不同的结果

大数据要想挖出有价值的信息，需从不同维度、全方位考虑。

2. 存储、计算、可视化

需要更高性价比的数据计算和存储方式。

3. 不同的数据管理策略

不同的数据管理策略即海量数据管理的问题。多数企业正在试图建设自己的数据中心，来满足大规模的数据量的产生。但随着数据的进一步增多，很多数据的查询和分析性能急剧下降，有的数据中心甚至出现了无法响应的状况，为企业的业务带来很大的损失。此外，还有数据的安全和隐私问题。

4. 超越企业现有 IT 的数据解决能力

只有少数企业（如跨国企业）有实力处理大数据，对中小企业而言，超越了本身的解决能力范围，加之无法招到大数据处理人才，所以，解决大数据问题变得非常困难。

（五）大数据处理技术

大数据处理技术包括数据采集（即从传感器、移动互联网、移动终端获取的数据）、对数据进行预处理、海量存储（数据存储很重要）、数据分析和挖掘、对结果进行可视化（方便决策）。

目前流行的大数据处理技术为 Hadoop 和 Spark。

Hadoop 的特点如下：

- 并行模式简单、编程较易。
- 为程序员屏蔽通信、并发、同步与一致性问题。
- 计算与存储一致，计算向数据靠拢，高效专用存储模式。
- 任务之间无依赖，具有高系统延展性。

Spark 是一种基于内存和 Hadoop 的分布式系统,克服 Hadoop 不擅长迭代计算的弱点,通常情况下比 Hadoop 快很多倍。

因此,大数据技术是真正的大智慧,可将大数据变废为宝。用户在线的每一次点击、每一次评论、每一个视频点播都是大数据的典型来源,通过互联网便可了解大数据带来的价值,如京东、阿里巴巴捕捉用户的消费行为,进而给出推荐。

(六) 大数据的隐私安全

大数据是一把“双刃剑”,在给企业带来财富的同时,往往也会对企业和个人带来伤害。

二、互联网+大数据

“互联网+”是利用互联网的平台,利用互联网技术、物联网技术、云计算与大数据技术等,利用互联网将某行业上下游及相关行业结合起来,在新的领域创造一种新的模式。创新是“互联网+”的灵魂,“互联网+”是创新的引擎。

大数据、“互联网+”在某一方面的创新,便是如何用一种模式吸引到用户,即用户的竞争,如 360 安全免费、机场免费 WiFi、滴滴打车等,均为大数据时代的商业模式。当然,有些基于资源的商业模式是不可复制的。

因此,互联网+大数据如何实现大智慧,取决于不同的行业应用,需要对具体行业进行具体分析,明确应用需求和行业痛点,才能建立一个有价值的行业大数据平台,产生真正的大智慧。

三、互联网+大数据=大智慧的案例

以阿里巴巴的信用贷款为例,阿里巴巴通过掌握的企业交易数据,借助大数据技术自动分析判定是否给予企业贷款,全程不会出现人工干预。目前阿里巴巴信用贷的坏账率为 0.3%左右,大大低于商业银行。大数据使阿里巴巴直接受益,成为整个阿里巴巴的财富。

大数据变废为宝的例子很多,如腾讯视频利用大数据推荐,迅速受到大

量用户的使用关注等，都为大数据成功的案例。目前，基于大数据的证券投资直接为投资者带来了帮助，如新浪、腾讯利用大数据对每家企业的信息进行收集，建立指数基金，从而帮助决策与投资。

四、创业实践——“我知盘中餐”大数据平台

创建平台的原因：源于我的三个梦想——帮助农民解决农产品销路问题；致力解决食品安全问题；理论联系实际，让大数据技术有用武之地。

为了实现这个梦想，创立了“我知盘中餐”互联网+大数据平台。为了实现大智慧，我们深入调研了餐饮产业链的痛点问题，如农产品销路、食品安全、餐饮企业成本高等。

痛点产生的原因：食材到达餐桌前会经过许多中间商，中间商需要赚钱，导致农民卖不出价钱，而城市消费价格又偏高。另外，农民想赚钱，就必须扩大生产，一扩大生产，产品又找不到销路。农民赚不到钱，更不注意产品质量，从而导致食品安全问题。

解决方法：即互联网+大数据，该平台是一个连接农产品供应商、餐饮企业、物流企业及广大消费者的餐饮采购产业链互联网+四大行业（现代农业、物流业、餐饮业、金融行业）平台，形成产业链的闭环。用户可以方便地在平台上进行采购和订餐，如图1所示。

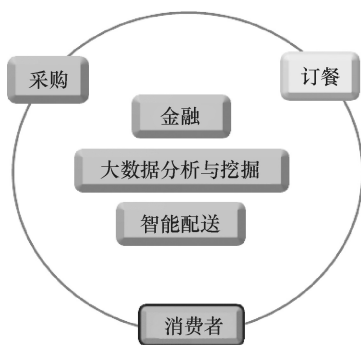


图1 “我知盘中餐”大数据平台

用大数据解决如下问题：

第一，农产品的销路问题。用大数据整合资源，实现按需生产和消费，形成订单式生产，精准扶贫，解决三农问题。

第二，吃的问题。用大数据实现农产品、餐饮企业质量排名推荐等，让大家采购到放心食材。

第三，食品安全问题。餐饮产业链大数据的有效整合，可以进行食品安全溯源，让造假者无处可藏。

第四，大物流问题。方便、快捷的智能配送系统，通过整合物流大数据，降低物流成本，同时缓解交通拥挤状况。

第五，智能对接、精准营销。让农业、餐饮、物流大数据插上科技的“翅膀”，实现精准营销，消费行为分析，客户关系管理，用大数据为用户创造价值。

关键创新点：核心技术创新为食品安全溯源、大数据、智能配送、人工智能。其他包括互联网金融、产业链整合等模式上的创新。

“我知盘中餐”平台对能保证品质的农产品供应商、餐馆、物流企业开通免费入驻绿色通道。通过消费者的采购和点餐，产品数据、餐饮数据、物流数据及消费者在平台留下的消费和点评数据等会越来越多。我的博士、研究生们对这些大数据进行深入研究和分析，从而创新出更多的产品，并始终保持技术的先进性，从而挖掘出更多有价值的信息，体现真正的大智慧，最终实现农餐大数据的共创、共享和共赢。

作者简介

张德富：博士（后）、教授，博士生导师，厦门大学计算机系副主任。闽江科学传播学者、厦门市科技经济促进会高级顾问。厦大大数据与计算智能团队带头人。研究方向包括大数据与云计算、计算智能、数据挖掘等，发表了40多篇SCI论文。承担国家自然科学基金、国家留学基金、国家社科重点基金大数据子课题，以及华为等企业大数据项目多项。创立我知盘中餐（厦门）电子商务有限公司。入选2017年厦门“双百计划”领军型创业人才，入

选 2017 年中英创新领军人才。曾获 2009 年福建省第六届高等教育教学成果奖一等奖（排名第 2）。指导厦大学生参加 ACM 世界大学生程序设计竞赛，获得 3 枚金牌、8 枚银牌，进入世界总决赛一次；指导厦大学生获得 2014 年全国 Spark 开发者竞赛桂冠。

在自由与控制之间达至创新

哈尔滨理工大学软件学院副院长 李 鹏

我们正处于信息时代，此前经历了农业时代、工业时代、电力时代，未来有人认为会是数据时代或智能时代。以计算机和互联网技术为基础的第三次工业革命确实产生了巨大的生产力，推动了社会的发展，也改变了我们的生活。

如今，互联网已经进入了社会各个领域的方方面面。数据如同水、电、石油、空气一样，成为一种资源。假如家里断网，就和停电停水一样难受，其实断得不是网，而是数据，难受的是对自由的限制。因为互联网就是一种自由的释放，数据与技术的发展使得我们更加自由。而大数据、人工智能、虚拟现实、物联网等技术的蓬勃发展，似乎也预示着我们能以令人震惊的速度步入下一个时代，整个世界似乎都疯狂了，当我们从这种亢奋的状态抽离出一些清醒时，我们在推动数据发展与技术共享、享受自由的同时，似乎忽略了一些东西，那便是控制与管理。数据与技术的不断进步，真的可以解决所有问题吗？真的能不断推动社会的发展吗？这是值得思考的问题。

作为一名高校教育工作者，当我们提供了近乎无限的学习资源，提供了近乎完美的学习平台，提供了最新的教学模式与方法，提供了个性化学习的指导，提供了超好的学习辅助工具，提供了优质的师资，提供了一切想要的自由，是否便可以培养出超出以往任何时代的优秀人才？

很多学校正在以飞一般的速度加快建设，力争达到上述标准。与过去的大学相比，学习的环境、资源、平台、技术支撑有着天壤之别，甚至没有可比性。很多东西在以前的时代是一种奢望，甚至无法想象。从学校的发展情况可以看出时代在不断进步、不断发展。但是，现在我们是否培养出了超出以往任何时代的优秀人才了呢？答案是没有。甚至“我的英语不如我老师，

我的学生英语不如我；我的编程不如我老师，我学生的编程不如我”成为一种现状。

1996 年的电脑远不如现在一部廉价手机的性能，但我们却用它撬开了编程世界的大门。当时我们为得到一本编程秘籍废寝忘食、争相传阅，为解决不了的问题相互争论、面红耳赤。记得当时在批发市场买得最多的是蜡烛和廉价电池，用来熬夜看书和听英语。而现在，遇到解决不了的问题搜索一下便可解决，MP3 也成为古董，编程资源、英语资源应有尽有，而学生的编程能力和英语水平却未提高。这为什么呢？是教育模式的失败、教育者的不尽心，还是学生的问题？其实都不是，而是人性使然。

当资源无限丰富，自由唾手可得时，我们失去了动力，我们失去了珍惜，我们失去了更高的人生追求（甚至理想）。刚才提到，数据如同煤炭、石油、电一样成为一种资源。在过去的 200 年，煤炭和石油为人类的发展与自由贡献了太多。现在煤炭还可使用 200 年左右，石油可以使用 50~100 年。我们意识到，除了付出巨大的环境代价外，人类似乎并没有为没有石油和煤炭做好技术上的准备。有人或许会说，数据资源不同，它不会枯竭，只会越来越多。但我们强调的不是资源的枯竭，而是资源的滥用。

从 1950 年伟大的图灵测试开始，人工智能的发展还不到百年。1995 年，“深蓝”击败卡斯帕罗夫，有人认为国际象棋并不算什么。但仅仅用了 20 年的时间，AlphaGo 先后击败了李世石和柯洁，从此宣布在棋类游戏上，人类对战人工智能失败。击败人工智能的唯一方法就是断电，不到百年的时间里自学习、机器学习、深度挖掘、神经网络……人类似乎在人工智能方面无比狂热、充满智慧，下一个百年甚至千年会有多少领域从此被人工智能击败无法得知。当数据无限扩大、技术不断进步，却没有相应的控制和管理水平，有一天人工智能或许会对人类这样说：你们太笨了，从呱呱坠地到学富五车需要那么多年，生命却那么短，比起我们自学习的速度和使用寿命，你们太脆弱了。或许目前看来还只是个玩笑，可当我们发现石油后无止境的利用，到今天我們也许做错了，可是来不及了。我并不是一个悲观主义者，只是要提醒人们在疯狂追求数据与技术自由发展的同时，也应该在控制与管理方面进行冷静的审视与研究。自由一定要被插上可以被控制的“翅膀”，否则，自由飞向的未必是天堂，也可能会是地狱。

作者简介

李鹏：哈尔滨理工大学软件与微电子学院副院长、教授。作为项目负责人，共承担国家自然科学基金、教育部、科技部、中国博士后、省、市等各级计划项目十余项，承担企业横向课题 6 项，累计科研经费 800 余万元；已发表 EI、SCI 检索论文 20 余篇，已授权专利 7 项；主要研究方向：网络信息处理、人工智能、自然语言处理等。荣获“校青年拔尖创新人才”“省高校青年学术骨干”等荣誉称号。

大数据安全机遇与挑战

河北师范大学信息技术学院院长 赵冬梅

一、大数据时代背景

随着信息技术的不断发展，以及云计算、物联网、社交网络等新兴技术和服务的不断涌现及广泛应用，数据种类日益增多，数据规模也呈现急剧增长的趋势，在这种形势下可以说大数据时代已经到来。据不完全统计，至 2016 年，全球信息量已达到 25ZB，数据量相当于 2.5 万亿 GB。到 2020 年，全球数据总量将会超过 40ZB，这个数据量是 2011 年的 22 倍。在过去几年，全球数据量大约以每年 58% 的速度增长，在未来这个速度会更快。如何利用大数据解决科学、医疗、能源、商业、政府管理、城市建设等领域的问题是摆在全世界面前的一大难题。《2015—2020 年中国大数据市场现状研究分析与发展前景预测报告》显示，未来中国大数据产品潜在的市场规模有望达到 1.57 万亿元，大数据的主要市场会集中于各实体企业对海量数据的处理和挖掘，而这些应用必然会带动数据存储设备、提供解决方案，以及大数据的分析、挖掘和加工类企业的市场。这一爆炸性的发展趋势同时会带来一个问题，即安全问题。

二、大数据安全面临的挑战

随着大数据的广泛应用，国家的大数据面临着国内外各种安全因素的威胁，这些关系到国民经济运行、社会政治稳定、国家安全利益的数据资源，一旦被国内外敌对势力利用，将会造成数据的流失、篡改和破坏，这就意味着国家的数据主权被侵犯，国家的安全出现漏洞。

（一）大数据安全面临的挑战

在我国，大数据安全面临着严峻的挑战，主要有以下几个方面。

第一，网络基础设施及基础软硬件受制于人。服务器、数据库等相关产品国外垄断严重。

第二，网站及应用漏洞、后门等不断出现。据统计，我国高达 60%的网站存在安全漏洞及后门，而我国各类大数据行业应用广泛采用各种第三方数据库、中间件，广泛存在漏洞。

第三，网络攻击手段更加丰富。其中终端恶意软件、恶意代码是黑客或敌对势力攻击大数据平台、窃取数据的主要手段之一。目前网络攻击越来越多地从终端发起，终端渗透攻击成为国家间网络空间安全战的主要方式。针对大数据平台的高级持续性威胁攻击也是很常见的。

（二）大数据安全研究的局限性和不足

目前，随着大数据时代的到来，大数据的安全研究存在局限性和不足，主要表现在以下几个方面。

第一，目前对攻击动机预测不足。因为在大数据环境下，针对数据资源所发起的攻击是在动机驱使下的多重组合、不确定和持续的攻击。如 2014 年的黑客入侵索尼影业等。目前的研究还需对攻击者实施攻击动机，即最终要达到的目标进行预测和分析。

第二，缺乏以大数据环境下数据保护为目标的研究。如今的网络安全研究大多以节点或局部网络的安全为评估目标，但在大数据环境下网络结构是动态的、不确定的，在这样的网络结构下，在数据资源开放、共享的前提下，以节点、局部网络安全为前提的研究便不适用于大数据环境下的数据保护。

第三，缺乏有效的网络攻防对抗结果分析。大数据资源及对应的网络安全防御设施面临着来自各方的直接或间接的攻击威胁，网络攻防双方处于互相对抗的过程。在攻击和防御的过程中，双方一方面是互有损失的，有可能退出；另一方面也是互有补充的。因此，对大数据环境下的网络攻防过程及结果进行有效分析，可以预测网络在大规模、持续攻击下的生存能力和数据资产保护的结果，以调整和合理配置网络防御资源。

通过以上几个方面可以发现，传统的网络安全研究面临巨大的挑战，各种安全设备报警、日志信息种类繁多，数据量大，为准确提取攻击意图增加了难度，数据存储和运行方式的改变使得目前的网络安全面临巨大的挑战。

三、大数据安全分析的机遇

（一）大数据的特征——“4V”“1C”

Variety: 支持多种类型的数据格式。

Volume: 大数据量的存储。

Velocity: 快速处理。

Value: 低价值密度。

Complexity: 指大数据的复杂性加大，同时提升了分析和处理大数据的难度。

（二）大数据是一把双刃剑

对安全问题而言，大数据是一把双刃剑，其结果取决于技术的使用者和目的。大数据的安全问题是其自身的对抗与博弈，是其自身固有的特点。其中涉及两个概念，一个是大数据自身的安全，包括针对大数据计算和大数据存储的安全性；另一个是基于大数据技术的安全，是指利用大数据技术来进行安全分析。因此，双刃剑也表现为两方面：一方面，因为大数据时代已到来，攻击方会利用大数据的特点对数据资源进行攻击，会影响大数据自身的安全；另一方面，对抗方也可以利用大数据技术进行防御。

随着大数据时代的到来，大数据的“4V”+“1C”特征为网络安全提供了全面的信息支持。具体表现如下：可以获取更多类型的日志数据，大数据分析的关联分析可以通过采用恰当的分析模型发现未知威胁，引入大数据分析技术可对若干年的数据进行分析。因此，威胁发现能力更强，寻找潜在的安全威胁对未发生的攻击进行防御，并对 APT 类供给进行有效应对。因此，大数据时代的到来为网络安全研究带来了机遇。

（三）大数据安全分析研究方向

借助大数据安全分析技术可以更好地解决海量安全要素信息的采集、存储问题，借助如机器学习、数据挖掘的一些算法，可以更加智能地洞悉信息与网络安全的态势，更加主动、弹性地应对新型复杂的威胁和未知多变的风险。主要的研究可以从如下两方面进行：

一方面是基于大数据技术的攻击意图预测。第一是主要集中于从网络自身结构的特点判断攻击意图，利用网络中的漏洞对路径进行预测，从而实现对攻击意图的识别，利用基础设施的相互依赖性识别攻击意图；第二是基于入侵检测信息，以攻击事件的数量、攻击路径为依据，运用概率统计等方法判断攻击意图；第三是基于博弈、马尔科夫链等思想判断攻击意图，基于博弈思想的识别方法主要是以博弈双方的策略和收益为准来推断攻击意图，马尔科夫链是依据入侵检测设备提供的报警信息对攻击意图进行预测。

随着大数据时代的到来及大数据分析的一些技术的出现，可以依托大数据对攻击者的历史信息进行全面的分析和提取，通过信息的提取可以获取、预测攻击意图，为安全态势的评估提供准确、全面的分析依据。

另一方面是基于攻防对抗的网络安全预测。在网络中，一方面网络攻击方通过直接或间接的方式对网络实施攻击。另一方面网络防御者一般是利用杀毒软件、防火墙、主动攻击等手段对网络攻击者进行防御，对抗之中双方互有损耗，同时，灾备系统也会对网络的恢复、运营起到相应的作用。因此，对网络安全的研究要从网络攻防对抗入手，才可以一方面评估网络安全态势，另一方面对网络中的安全设施配置给出建设和指导意见，才能保证网络在不同类型的攻击者、不同持续时间、不同程度的攻击条件下的安全态势及防护措施的有效性。

通过大数据的安全研究可以更好地解决海量数据安全要素信息的采集、存储问题，借助大数据分析技术的机器学习和数据挖掘算法，更加智能地感知信息与网络安全的态势，主动、弹性地应对新型复杂的威胁和未知多变的风险。

作者简介

赵冬梅：西安电子科技大学计算机应用专业博士研究生学历，博士学位。现为河北师范大学教授，学科带头人，硕士生导师，河北省信息化专家委员会委员。在网络信息安全风险评估领域取得了多项研究成果。近年来发表学术论文 20 余篇。主持或承担了教育部科学技术重点项目 1 项；河北省自然科学基金 2 项；河北省科技厅项目 4 项；河北省教育厅基金项目 3 项；主持的项目获得河北省科技进步奖三等奖 1 项。

后 记

市场上大数据图书名目很多，但多数图书主要介绍大数据概念、技术、管理和应用等方面的内容，与学校大数据教育教学的专业内容关联不大。相对来说，《赋能大数据教育：全国高校大数据教育教学经验谈》这本书，内容比较专业、专注。一是聚焦高校，二是聚焦老师。书中内容全部由高校在职老师结合本校大数据教育教学实践经验及愿景规划进行分享。思想新，干货多，接地气。

大数据其实并不神秘，大数据时代的到来，让社会科学领域的发展和研究从宏观群体逐渐走向微观个体，让追踪每一个人的数据成为可能，从而让研究每一个个体成为可能。对于教育研究者来说，我们将比任何时候都更接近发现真正的学生，而这，正是教育的进步。

大数据对于教育的改变将会是补充，而不是颠覆！

出版《赋能大数据教育：全国高校大数据教育教学经验谈》这本书，是教育变革的需要。大数据已经上升为国家战略，我们已经进入大数据时代，大数据发展前景毋庸置疑，大数据产业发展要从教育行业抓起，《赋能大数据教育：全国高校大数据教育教学经验谈》这本书能够开阔高校师生及教育界的教学视野，在大数据教育教学中，能够启发思想，取长补短，借鉴经验。

出版《赋能大数据教育：全国高校大数据教育教学经验谈》这本书，是时代市场的需要。教育不只是“你讲我听”、考试评分或是选修科目而已。历史上第一次，我们拥有了强大的、具有实证效果的大数据工具，能够空前地看到学习的过程，破解过去不可能发现的重重学习阻碍，让教育可以实现“私人定制”，改善学习的成效。教师的工作不但不会被网络视频所代替，还会变得更高效、更有趣，学校和政府部门也能用更低的成本提供更多的教育机会。在这一刻，我们可以清晰地看到：一个全新的大数据教育时代正在到来！

出版《赋能大数据教育：全国高校大数据教育教学经验谈》这本书，是高校发展的需要。旨在向教育界及全国高校师生普及大数据知识，推广大数据应用。截至目前，我国有 35 所高校已经开办“数据科学与大数据技术”等相关专业。2017 年又有 263 所本科院校向教育部申报“数据科学与大数据技

术”专业，其中工学 190 所，理学 73 所。作为交叉型学科，大数据专业的相关课程涉及数学、统计和计算机等学科知识，“数据科学与大数据技术”专业也强调培养具有多学科交叉能力的大数据人才。《赋能大数据教育：全国高校大数据教育教学经验谈》这本书，第一次向全国高校师生展示了当前我国大数据教育教学现状、愿景，以及问题和解决方案。

出版《赋能大数据教育：全国高校大数据教育教学经验谈》这本书，是人才培养的需要。目前，迫切需要解决的是大数据人才培养问题。相关调查显示，未来 3~5 年，我国大数据人才缺口达到 150 万之多。关于大数据人才培养及数据科学专业研究工作，我国教育界似乎远未做好准备。据了解，直到目前，还有许多高校计算机、软件专业依然在使用 20 世纪 90 年代初的落后教材。目前高校正在使用的计算机课程体系、教学内容，整体上已经落后于当前流行的 IT 技术潮流 3~5 年。《赋能大数据教育：全国高校大数据教育教学经验谈》这本书，能够促使全国高校师生对以大数据、人工智能等为代表的新一代 IT 技术有一个全面的了解、全新的认识，对普及新一代 IT 技术和知识起到重要的桥梁纽带作用。

未来，以大数据、人工智能等为代表的新一代 IT 技术将彻底解放一些具有创新精神的老师，能使他们抛弃大量重复的劳动而将精力集中在教学的核心功能，这就是技术的解放力量。

在本书出版过程中，得到了全国 41 所高校 51 位老师的大力支持，他（她）们为本书倾注了大量时间和心血。另外，还要非常感谢电子工业出版社学术出版分社的大力支持，社长董亚峰先生、编辑缪晓红女士对全部书稿内容进行反复修改，保证了本书的质量。还要感谢 CIO 时代 APP 的编辑人员，侯丽敏、孔文两位编辑在此书编辑出版过程中全程对接、服务 51 位作者，工作有条不紊，一丝不苟，经常加班到深夜，为本书的出版付出了大量时间和精力，在此，一并表示感谢！

朱启明
全国高校大数据教育联盟 副秘书长
CIO 时代 APP 总编辑

2017 年 12 月

